

## **CAPÍTULO II.4**

# **APLICACIONES DEL SGML A LA DOCUMENTACIÓN JURÍDICA**

## II.4.1. INTRODUCCIÓN

La difusión y los intercambios de información desempeñan un papel central en la vida de la comunidad legal, tal y como se ha venido argumentando a lo largo de esta tesis. Por este motivo, la aplicación de estándares documentales, como es el SGML, en los ámbitos de la actividad legislativa y judicial, será uno de los pilares básicos en los que se han de sustentar todos los procesos informativos que se producen en esta esfera de elaboración, gestión y difusión de documentos electrónicos de carácter legal. La aplicación del SGML dentro de este entorno ha supuesto, según palabras de D. Poulin, una profunda transformación en la forma de concebir y utilizar la documentación, así como en la manera en que se producen y conducen los intercambios de información dentro de esta comunidad<sup>171</sup>. Así, la adopción de este estándar por parte de editores, cortes legislativas, tribunales de justicia, bufetes de abogados, etc., supone aprovecharse de todas las ventajas que llevan asociadas las aplicaciones SGML para la producción, gestión, recuperación y reutilización de documentos electrónicos, tratado todo ello en los capítulos previos.

Bien es cierto que dentro de la comunidad profesional de los especialistas en el tratamiento de la documentación jurídica ha habido opiniones diversas sobre la idoneidad de unos estándares u otros para la edición electrónica de este tipo de documentos. Propuestas que han ido desde la aplicación del lenguaje TeX, adaptado a las necesidades especiales de los documentos jurídicos dando lugar así a una versión de dicho lenguaje conocido como LeXTeX<sup>172</sup>, hasta otras propuestas más extendidas, como la adaptación del estándar EDI (*Electronic Data Interchange*) para llevar a cabo ciertas funciones de distribución de información electrónica dentro de las instituciones de carácter jurídico. Pero a pesar de

---

<sup>171</sup> Daniel Poulin. "Le SGML et son intérêt pour la gestion des documents juridiques" [documento HTML]. *Lex Electrónica*, v. 2, n° 3, automne 1996. Disponible en <http://www.lex-electronica/articles/v2-3/poulinfr.html> (consultado el 22 de noviembre de 2000).

<sup>172</sup> Roger Horne. *Markup, Typesetting and the Law* [documento HTML]. The Roger Horne's Miscellany, [sin fecha]. Disponible en <http://www.number7.demon.co.uk/lextex/lextex.htm> (consultado el 8 de enero de 2001).

esta disyuntiva, el estándar SGML ha venido ganando la batalla en todo este tiempo, como se explicará más adelante.

Como ya se expuso, SGML fue concebido en su origen para adaptarse a cualquier ámbito de producción y gestión documental y, por tanto, la esfera constituida por los documentos legales, con características inherentes tan propicias para ser tratadas de forma estructural, supone uno de los claros ejemplos de aplicación de este lenguaje basado en el marcado descriptivo y en la definición de tipos documentales. Tan es así, que la documentación jurídica, y el texto en ella contenido, presentan las características típicas de cualquier aplicación SGML; a saber<sup>173</sup>:

- Intercambio: circulación de los mismos textos entre los actores que pueden hacer uso de esta documentación (legisladores, jueces, abogados, secretarios, editores especializados, etc.).
- Complejidad: gran variedad de tipos documentales (leyes, reglamentos, jurisprudencia, manuales y artículos doctrinales, manuales de interpretación y aplicación de leyes, etc.) con múltiples referencias cruzadas entre todos estos documentos.
- Calidad: los documentos jurídicos han de ser redactados con una calidad extrema al no permitirse errores en los textos contenidos.
- Reciclado: los textos jurídicos pueden ser reciclados para la elaboración de otros productos documentales (proyectos de ley, diarios oficiales, leyes refundidas, boletines jurídicos de índices y sumarios, bases de datos, etc.).
- Reutilización: similar a la anterior, en donde se reutilizan estos textos para enmiendas, anotaciones legislativas, manuales de aplicación, publicaciones comerciales, etc.
- Longevidad: los documentos jurídicos suelen tener una vigencia prolongada y en muchos casos es necesario seguir la evolución histórica que se ha producido desde su redacción inicial.

---

<sup>173</sup> Marc-André Ledoux. *SGML au service de la recherche documentaire juridique* [documento HTML]. Montreal: Université de Montreal, Faculté de Droit, Centre de Recherche en Droit Public, 8 de julio de 1996. Disponible en <http://www.lexum.umontreal.ca/fr/equipes/technologie/conferences/sgmlquebec/10.html> (consultado el 17 de noviembre de 2000).

- Gestión de calidad: los poderes y administraciones públicas están obligados a realizar todos sus procesos de gestión con la mayor calidad posible.
- Adaptación tecnológica: resulta imperativa la difusión de estos documentos en los nuevos soportes tecnológicos que se vayan introduciendo (ópticos, Internet, intranet, etc.).

Las aplicaciones que del lenguaje SGML se pueden integrar dentro de los sistemas informáticos para el tratamiento de la documentación jurídica son múltiples y diversas, destacando entre ellas, según M. Blaquier, las siguientes<sup>174</sup>:

#### 1. Edición asistida por ordenador de documentos jurídicos:

Se trata de la labor inicial para la redacción de textos jurídicos, utilizando para ello un editor de textos capaz de elaborar y manipular textos electrónicos y marcas SGML. Además de las funciones habituales de un editor de textos, este tipo de programas debe proporcionar una serie de funcionalidades, entre las que se pueden destacar las del reconocimiento de documentos SGML, generación, gestión y actualización de DTDs, importación y exportación de documentos SGML, conversión de documentos electrónicos marcados con otros formatos (públicos o propietarios) a documentos SGML, etc. Estos editores deben ser capaces, asimismo, de generar documentos SGML de acuerdo a una DTD definida y establecida previamente para cada uno de los posibles tipos documentales que se manejan en la institución. Las restricciones que se establecen en cada una de las DTDs a la hora de elaborar los documentos electrónicos asegurarán la consistencia y validez de los documentos generados.

Existen en el mercado numerosos productos informáticos de este tipo con diversos niveles de funcionalidad o especialización, aunque muchos de ellos, lógicamente, se han adaptado o reorientado en la actualidad al trabajo con documentos XML. Destacar por su uso e implantación programas de la talla de *Mark-it* de Sema Software Technology, el

---

<sup>174</sup> Marta Blaquier Ascaño. “Aplicaciones del SGML a la información jurídica”. *Informática y Derecho*, nº 12-15, 1996, p. 1490.

*Author/Editor* de SoftQuad (integrado dentro del paquete *Panorama*), *FrameMaker+SGML* de Adobe Systems, ADEPT de la compañía ArborText, *SigmaLink* de la compañía alemana STEP (*Stürtz Electronic Publishing GmbH*), el programa *Near & Far* de Microstar para la modelización y desarrollo de DTDs o, por último, adaptaciones para el trabajo con textos SGML de procesadores de texto tradicionales, como es el caso de *WordPerfect*<sup>175</sup>.

## 2. Recuperación y procesamiento de información jurídica:

La generación de documentos electrónicos según el lenguaje SGML tiene, como se ha venido indicando anteriormente, un enorme potencial para la búsqueda y recuperación de información, y en nuestro caso, de interés para el jurista (y en ocasiones, sobre todo en el caso de la legislación, para el común de la ciudadanía). Los documentos SGML facilitan enormemente la generación de sistemas informáticos con potentes funcionalidades de interrogación y localización de la información demandada, bien sea a través de la generación de bases de datos documentales o relacionales de interés jurídico, o bien a través de sistemas de recuperación directa de documentos. Además, las capacidades hipertextuales que proporciona el estándar asociado HyTime permite aumentar estas capacidades dado que nos permite realizar una consulta y lectura no secuencial del texto a través de la “navegación” de las relaciones establecidas entre los documentos. Este aspecto es de vital importancia en documentación jurídica debido a que la lectura de los textos jurídicos por parte de los profesionales del Derecho, como acertadamente expone A. Gonzales Aguilar, nunca suele ser secuencial: “él trabaja o analiza un documento, interrumpe su lectura para ir a consultar una ley, o mejor un artículo o algunos artículos precisos de una ley, o una jurisprudencia”<sup>176</sup>. Los sistemas de hipertexto se revelan, como veremos en capítulos posteriores de esta tesis, como una de

---

<sup>175</sup> En cualquier caso, para profundizar sobre este tema volvemos a recomendar el directorio de herramientas informáticas SGML mantenido por Steve Pepper y disponible en <http://www.infotek.no/sgmltool/>

<sup>176</sup> Audilio Gonzales Aguilar. “Hipertexto y derecho”. *Informática y Derecho*, n° 5, 1994, p. 858.

las soluciones idóneas para el tratamiento de los documentos jurídicos debido a la gran cantidad de relaciones que se pueden establecer entre ellos.

### 3. Transmisión e intercambio de información electrónica de interés jurídico:

La transmisión e intercambio de información a través de medios electrónicos viene jugando desde hace ya varias décadas un papel crucial para el desarrollo y avance de cualquier organización, no siendo una excepción las instituciones legales y jurisprudenciales. Es cierto que para esta función se han desarrollado otros importantes estándares internacionales, entre los que destaca el caso del EDI (*Electronic Data Interchange*), el cual fue extendido para adaptarse a las necesidades de almacenamiento, gestión, localización e intercambio de información electrónica dentro de los tribunales de justicia, dando lugar, por poner el caso de la justicia norteamericana, al llamado JEDDI (*Judicial Electronic Document and Data Interchange*)<sup>177</sup>. Según datos aportados por J. E. McMillan, Director del Court Technology Laboratory del National Center for State Courts de los Estados Unidos, en este país se generaban a principio de la década de los 90 más de 25 millones de documentos fruto de la actividad diaria de la judicatura estatal y federal (si se cuenta aquí la documentación que generan las apelaciones a las decisiones de estos tribunales, los documentos administrativos y los tribunales de jurisdicción inferior la cifra ascendería a los 83 millones de documentos por año)<sup>178</sup>. En este contexto, el uso de la aplicación JEDDI se basaba en la elaboración normalizada de

---

<sup>177</sup> El NCSC (*National Center for State Courts*) de los Estados Unidos de América apostó fuertemente a principios de la década de los 90 por este estándar como modelo para el tratamiento electrónico de la información judicial que se produce en los diferentes tribunales de justicia del estado. Las razones esgrimidas se apoyaban en las ventajas aportaba en aquel momento: ser un estándar independiente del hardware y del software para el intercambio de datos electrónicos, el proporcionar un mecanismo seguro para la autenticación legal de los documentos transmitidos, reducir los costes asociados dentro de los tribunales de justicia y facilitar la comunicación y transmisión de información judicial entre los tribunales de menor rango y los de apelación. Para una más amplia información sobre los argumentos defendidos a favor de este estándar véase *The Need for National Standards for Judicial EDI* [documento HTML]. Williamsburg, VA: National Center for State Court, 21 de junio de 1991. Disponible en <http://www.ncsc.dni.us/jeddi/edijust.htm> (consultado el 8 de enero de 2001). En cualquier caso, toda la información oficial relativa a JEDDI puede ser consultada en el sitio Web de la institución mencionada en la dirección <http://www.ncsc.dni.us/jeddi/jeddi.htm>

ficheros electrónicos con aquella metainformación (metadatos) contenida en los expedientes judiciales, asociando a los mismos las imágenes digitalizadas de los documentos integrantes de dichos expedientes, y siendo transferido todo ello a través de redes informáticas de transmisión de datos<sup>179</sup>.

SGML ha competido directamente con este estándar ganando en muchos casos la batalla, debido a que combina perfectamente la rigurosidad en la transmisión de datos (el mayor beneficio de JEDDI) con la potencialidad en el manejo de documentos electrónicos estructurados; se centra tanto en los datos como en los documentos electrónicos que los contienen<sup>180</sup>. Cuando se piensa en la utilización del SGML para el intercambio electrónico de información jurídica se emplean, normalmente, dos estrategias:

1. Es posible que varias instituciones que intercambien documentos electrónicos SGML hayan elaborado sus propias DTDs para cada uno de los tipos documentales que maneja. En este caso, será necesario un proceso complejo de armonización de estas DTDs.
2. Pero también es posible, y absolutamente recomendable, que existan definiciones de tipos documentales comunes para todas las instituciones que han de participar en los procesos de transferencia de información jurídica, dándose de este modo una total y completa compatibilidad documental en todo el proceso.

---

<sup>178</sup> James E. McMillan. *Judicial Electronic Document and Data Interchange (JEDDI): A Roadmap* [documento HTML]. Williamsburg, VA: National Center for State Court, February 22, 1994. Disponible en <http://www.ncsc.dni.us/jeddi/overview.htm> (consultado el 8 de enero de 2001).

<sup>179</sup> *Ibid.*, <http://www.ncsc.dni.us/jeddi/overview.htm>

<sup>180</sup> Daniel Poulin, Marc Tremblay, Serge Parisien. *Échange électronique d'informations juridiques; expériences récentes et perspectives* [documento HTML]. Montreal: Université de Montreal, Faculté de Droit, Centre de Recherche en Droit Public, [1995?]. Disponible en <http://www.lexum.umontreal.ca/fr/equipes/technologie/textes/sgml/echange.html> (consultado el 8 de enero de 2001).

Las experiencias más importantes e interesantes dentro de la aplicación de los sistemas SGML a la documentación jurídica se han producido desde esta segunda vertiente de trabajo cooperativo, como veremos posteriormente.

Destacaremos igualmente en este punto que SGML ha contribuido de forma trascendental en estas últimas décadas al intercambio de información electrónica de valor jurídico a escala mundial a través de una de sus más famosas aplicaciones, el HTML (*HyperText Markup Language*), como se vio en el último capítulo de la primera Parte de esta tesis doctoral al hablar de la actual presencia en Internet de recursos de información jurídica (apartado I.3.2). Este lenguaje, que será analizado en mayor profundidad en el siguiente capítulo, ha servido, y sirve, de importante soporte para la difusión e intercambio de documentos electrónicos jurídicos hipertextuales en el espacio virtual conocido como la *World Wide Web* dentro de Internet<sup>181</sup>.

Por último, apuntaremos que a pesar de los beneficios que se pueden derivar del empleo de SGML para la generación de documentos jurídicos electrónicos y de los numerosos ejemplos que se han venido dando en diversas instituciones jurídicas de todo el mundo, como se expondrá a continuación, la aplicación del SGML a la documentación jurídica no está exenta de problemas. Sin entrar en un profundo detalle sobre sus limitaciones, dado que el lenguaje de marcado de documentos electrónicos objeto del interés de esta tesis es XML, sí es necesario resaltar que en un informe elaborado por la Law School de la University of South California para la RLG Digital Collection Project Task Force<sup>182</sup> se hace mención a dos de los principales obstáculos con los que se enfrenta el lenguaje SGML para

---

<sup>181</sup> Marta Blaquier Ascaño. "SGML applications to legal information". *European Journal of Law, Philosophy and Computer Science*, v. 1, n° 2, 1995, p. 230.

<sup>182</sup> La RLG (*Research Library Group*) es una poderosa organización sin ánimo de lucro fundada en 1974 por las universidades americanas de Columbia, Yale y Harvard y la biblioteca pública de la ciudad de Nueva York con la misión principal de proporcionar acceso a la información necesaria para la investigación científica y el aprendizaje. En la actualidad se integran dentro de esta organización más de 160 universidades, bibliotecas nacionales, archivos y otro tipo de instituciones afines. Para una mayor información consúltense las páginas web de esta institución en <http://www.rlg.org/>

su implantación en todas aquellas instituciones que manejan documentación de valor jurídico; éstos son<sup>183</sup>:

- La ausencia de una Definición de Tipo Documental estandarizada y de alcance nacional o internacional para este tipo de documentos, a pesar del gran número de proyectos que se han venido desarrollando en algunas instituciones jurídicas de diversos países. Este hecho ha provocado que algunas instituciones hayan adaptado DTDs o aplicaciones SGML de propósito general y de gran alcance y reputación, como es el caso de TEI<sup>184</sup>, EDI (como se verá posteriormente) y, posteriormente, HTML, para el tratamiento de los textos legales.
- La gran complejidad de los sistemas documentales basados en SGML, tanto por las labores propias del marcado de los textos (labor tediosa y que acarrea un importante consumo de tiempo) como por el elevado precio de las herramientas informáticas necesarias para su procesamiento y gestión.

---

<sup>183</sup> *SGML in the Law* [documento HTML]. Los Angeles, CA: University of South California, Law School, 12 de septiembre de 1995. Disponible en <http://www.usc.edu/dept/law-lib/rlg/sgml.html> (consultado el 5 de octubre de 2000).

<sup>184</sup> Nicholas D. Finke. *TEI Extensions for Legal Text* [documento HTML]. Providence, Rhode Island: Brown University, Computing and Information Services, 1997. Disponible en <http://www.stg.brown.edu/conferences/tei/te10/tei10.papers/finke.html> (consultado el 23 de octubre de 2000).

## II.4.2. PRINCIPALES EXPERIENCIAS

La aplicación del lenguaje SGML para la construcción de documentos electrónicos en el ámbito jurídico ha sido paradójicamente tardía, pues no es hasta pasados algunos años de la década de los 90 cuando se puede hablar de la aparición de proyectos de envergadura para el tratamiento de documentos de carácter legal. Y decimos que este hecho resulta paradójico dado el origen del lenguaje SGML, pues recordemos que su predecesor, GML, fue inicialmente experimentado por Charles F. Goldfarb y su equipo de colaboradores para automatizar y gestionar los procesos documentales que se realizaban en una oficina judicial.

Como acertadamente señala M. Blaquier, las primeras aplicaciones del estándar SGML en el campo jurídico fueron ligadas a la publicación de documentos normativos por diversas instituciones editoras especializadas en la publicación de este tipo de información, tanto del sector privado como del público<sup>185</sup>, que emplearon inicialmente el SGML como medio idóneo para una posterior publicación de la información de carácter legal a través de los tradicionales medios impresos. Entre las empresas editoras privadas cabe destacar por su pronta utilización para la confección, almacenamiento y publicación de normas legales a la casa editora **Kluwer**<sup>186</sup> en Holanda y la **Thompson Legal Publishing**<sup>187</sup>, que creó el *Research Institute of America and Carswell* como subdivisión para la investigación y uso de SGML en la publicación de documentos jurídicos. Pero es sin duda en el sector de los editores públicos donde se han llevado a cabo las experiencias de trabajo con SGML más importantes. Resultan de especial interés las tres siguientes:

- La **Government Printing Office**<sup>188</sup> de los Estados Unidos anunció en 1995 a través de su editor jefe, Michael F. Dimario, la transferencia de los documentos producidos por

---

<sup>185</sup> M. Blaquier Ascaño. “Aplicaciones del SGML...”. *Op. cit.*, p. 1489.

<sup>186</sup> <http://www.kluwer.nl/>

<sup>187</sup> <http://www.thompson.com/>

<sup>188</sup> <http://www.access.gpo.gov/>

el Congreso americano a un sistema basado en el lenguaje SGML para, de este modo, facilitar las tareas de búsqueda y localización de información de carácter oficial<sup>189</sup>. La GPO, que fue creada con la finalidad principal de satisfacer las necesidades de material impreso del Congreso, es en la actualidad el punto focal para la impresión, recopilación y disseminación de información para toda la comunidad de estados federados. La gran experiencia que ha venido adquiriendo en estas últimas décadas en el tratamiento de documentos electrónicos de carácter legislativo y su publicación en diversos medios, le hace estar a la vanguardia de los suministradores de información gubernamental, empleando para ello un gran abanico de formatos de distribución de esta información (material impreso, microfilmes y microfichas, CD-ROM y redes informáticas).

- La **Her Majesty's Stationary Office**<sup>190</sup> es la imprenta y editora oficial del gobierno británico, encargada de la publicación de toda la normativa legal generada por el parlamento de esta nación. Esta institución fue una de las primeras usuarias de los sistemas de elaboración de textos asistidos por ordenador, llegando a desarrollar en 1972 su propio sistema de composición tipográfica basado en un lenguaje de codificación genérica<sup>191</sup>. Igualmente fue pionera en la elaboración de proyectos de envergadura para el desarrollo de bases de datos electrónicas de contenido jurídico. Se trata, además, de una de las primeras instituciones públicas de todo el mundo en hacer uso del lenguaje SGML, pues desde comienzos de 1987 toda la legislación emanada del parlamento británico era preparada y codificada en SGML de acuerdo a una compleja DTD, con la doble intencionalidad de que el marcado de los textos legales fuese de utilidad tanto para los sistemas de impresión y publicación como para el sistema de almacenamiento y gestión de la información en las correspondientes bases de datos, de tal forma que una nueva ley pudiese aparecer al mismo tiempo publicada en el boletín

---

<sup>189</sup> SGML in the Law. *Op. cit.* <http://www.usc.edu/dept/law-lib/rlg/sgml.html>

<sup>190</sup> <http://www.hmsso.gov.uk/>

<sup>191</sup> D. Barron. *Op. cit.* <http://cajun.cs.nott.ac.uk/compsci/epo/papers/volume2/issue1/epdx021.pdf>. Los lenguajes de codificación genérica se trataron en el capítulo II.1, apartado II.1.2.3.

impreso tradicional así como en formato electrónico<sup>192</sup>. A principios de 1996 se anunció que toda la información legal contenida en la HMSO sería puesta a disposición del público en la red Internet, sin coste alguno para el usuario por la consulta de la misma. Este pionero e importante desarrollo está basado en el proyecto denominado *The Statute Law Database*, versión electrónica del tradicional boletín impreso *Statutes in Force*. Esta base de datos sigue recogiendo la legislación producida desde los años 80 hasta nuestros días, marcada en SGML, para que sea consultada a través de Internet por cualquier usuario, proporcionando la visualización de los documentos obtenidos tras la consulta en formato HTML<sup>193</sup>.

- La **Official Publication Office**<sup>194</sup> de las Comunidades Europeas desarrolló a mediados de la década de los 80 un sistema denominado FORMEX (*Formalised Exchange of Electronic Documents*) con el fin de facilitar la producción de las publicaciones oficiales de la Comisión, en especial el Diario Oficial, así como la mejora en los sistemas de almacenamiento y recuperación de estos documentos en formato electrónico. Este sistema ha venido evolucionando notablemente desde su lanzamiento, con diversas versiones mejoradas, constituyendo hoy en día uno de los ejemplos más atractivos del empleo del SGML para el tratamiento y publicación de documentos de carácter legal<sup>195</sup>. La Unión Europea ha apostado fuertemente por SGML como estándar documental de trabajo, desarrollando múltiples DTDs para los diversos tipos documentales que se

---

<sup>192</sup> R. Stutely. "HMSO's Database of Legal Text". *SGML Users' Group Bulletin*, v. 2, n° 2, 1987, pp. 1987. Citado en David Barron. *Op. cit.*, <http://cajun.cs.nott.ac.uk/compsci/epo/papers/volume2/issue1/epdxb021.pdf>

<sup>193</sup> Roger Horne. *The Statute Law Database* [documento HTML]. The Roger Horne's Miscellany, 8 de abril de 1997. Disponible en <http://www.number7.demon.co.uk/papers/SLD/SLD.htm> (consultado el 8 de enero de 2000).

<sup>194</sup> <http://www.eur-op.eu.int/>

<sup>195</sup> Para una mayor información sobre este aspecto, véase la información oficial que suministra esta institución en <http://www.eur-op.eu.int/general/fr/b4.htm>. Existe igualmente una publicación oficial impresa sobre este desarrollo elaborada en 1985 por la propia OPOCE: Office for Official Publications of the European Communities. *FORMEX – Formalized Exchange of Electronic Publications*. Luxembourg: OPOCE, 1985.

producen en sus organismos dependientes. Dentro del ámbito jurídico es igualmente destacable por su importancia y uso extendido entre los profesionales del derecho comunitario el sistema automatizado de documentación para la información legal emanada de la UE, conocido como CELEX (*Comunitatis Europae LEX*)<sup>196</sup>. Esta magna base de datos, producida como un sistema interinstitucional, es alimentada a través del sistema de entrada de datos de diferentes fuentes, denominado ALICE (*ALIMENTation Celex*), que está basado por completo en el estándar SGML, utilizado no sólo para la definición del formato de entrada de los documentos sino también para el formato básico bajo el cual los documentos son almacenados para el intercambio entre el sistema de producción y de difusión de las distintas bases de datos<sup>197</sup>.

En el ámbito de la administración de justicia también encontramos notables ejemplos de aplicación del estándar SGML al tratamiento de la documentación producida. La incorporación de los tribunales de justicia al tratamiento y procesamiento de información electrónica en los países más desarrollados del mundo vino inicialmente de la aplicación de otros estándares para el intercambio electrónico de datos, como era el caso del EDI aplicado a este tipo especial de información, como ya se comentó con anterioridad (apartado II.4.1 de este capítulo). Algunos de los proyectos más destacables en este campo los podemos encontrar en Estados Unidos y Canadá.

Para el caso de los **Estados Unidos**, encontramos ejemplos como el del **Delaware Superior Court** con el lanzamiento del proyecto CLAD (*Complex Litigation Automated Docketing*)<sup>198</sup> con la finalidad de facilitar la gestión de litigios de gran complejidad que se

---

<sup>196</sup> Toda la información referente a esta importante base de datos, así como su consulta, es accesible en la dirección <http://europa.eu.int/celex/>

<sup>197</sup> J. Marín Navarro, P. E. Alevantis. "Alice in the Wonderland of SGML: streamling text entry in the CELEX databases". *The Electronic Library*, v. 9, n° 3, June 1991, p. 155.

<sup>198</sup> Una profunda descripción histórica y funcional de este importante proyecto se puede encontrar en la comunicación de Robert J. Katzenstein. "CLAD – Delaware's Paperless Docket" [documento HTML]. En: *National Court Technology Conference (CTCA)* (4°. 1994. Washington D.C.). Washington D.C.: National Center

producían en dicho tribunal estatal, demostrándose la viabilidad de esta tecnología y la gran reducción de costes económicos que llevaba aparejada.

Igualmente interesante resulta el proyecto del **Snake River Basin Adjudication (SRBA) – District Court** del estado de Idaho<sup>199</sup>, que surge de forma parecida al anterior, esto es, para dar respuesta a una situación jurídica de gran complejidad: la gran cantidad de reclamaciones tramitadas en este tribunal debido a la repartición de las aguas de la cuenca del río Snake. La administración de este tribunal buscó la mejor fórmula que permitiese automatizar la gestión de los expedientes producidos, encontrándola en la utilización de formularios electrónicos expedidos desde ordenadores personales y utilizando las líneas telefónicas para su transmisión.

Pero sin duda, el proyecto más ambicioso le corresponde a la **Utah Administrative Office of the Courts**<sup>200</sup>, citada numerosas veces como el caso más ejemplar de utilización del SGML en el campo de la administración de justicia norteamericana. Los tribunales de justicia de este estado americano han estado investigando durante las pasadas décadas las posibilidades para el intercambio de información jurídica de estándares como EDI (a través del desarrollo JEDDI, explicado con anterioridad) y el SGML, desarrollando diversos proyectos al respecto pero decantándose principalmente por la segunda opción. Este hecho se debió, según de A. Asay, a que los documentos son el centro de atención en cualquier caso de litigio y el estándar EDI se centra principalmente en los datos para la construcción de bases de datos e intercambio de los mismos, ignorando al documento en sí mismo, cosa que no ocurre, como ya se ha expuesto, con el estándar SGML<sup>201</sup>. De hecho, a este abogado e investigador en las tecnologías de la información aplicadas al derecho, y al propio Goldfarb (igualmente jurista), se les considera los primeros visionarios en la

---

for State Courts, October 1994. Disponible en [http://www.ncsc.dni.us/ncsc/ctc4/articles/elec\\_b.htm](http://www.ncsc.dni.us/ncsc/ctc4/articles/elec_b.htm) (consultado el 25 de noviembre de 2000).

<sup>199</sup> Toda la información oficial relativa a este proyecto se encuentra disponible en <http://www.idwr.state.id.us/info/water/srba/main.htm>

<sup>200</sup> <http://courtlink.utcourts.gov/>

<sup>201</sup> Alan Asay. *Applying Information Technology to Court Case Files: Report of the Utah Electronic Filing Project* [documento HTML]. Salt Lake City, UT: Utah Administrative Office of the Courts, septiembre de 1994. Disponible en <http://www.law.indiana.edu/cphomepg/efiling1.html> (consultado el 8 de enero de 2001).

aplicación de los lenguajes de marcado de texto a los documentos electrónicos de carácter jurídico. Tan es así, que en 1987 (un año después de ser declarado SGML estándar internacional), Asay desarrollaría las primeras DTDs SGML aplicadas al campo del derecho civil y criminal de aplicación en los tribunales de justicia del estado de Utah. A partir de estas primeras experiencias, los tribunales de este estado norteamericano se embarcaron en un importante proyecto piloto, conocido por el nombre de *Utah Electronic Filing Project*, de utilización de tecnologías informáticas para agilizar todos los procesos administrativos y judiciales que se dan en estas instituciones, donde el estándar SGML ha jugado un papel crucial en el desarrollo de este servicio. El sistema posibilita la recogida de los documentos necesarios para los litigios a través del correo electrónico, actualizando de forma automática la base de datos del tribunal (así como los sistemas de contaduría) y almacenando todos los documentos necesarios en formato electrónico. Los documentos propios generados por los tribunales de Utah son marcados a través del lenguaje SGML, ajustándose a una serie de definiciones de tipos documentales desarrolladas *ad hoc* y definidas en una guía de uso interno, denominada *Practitioner's Guide to Electronic Filing in the Utah Courts*.

Destacaremos igualmente en los Estados Unidos de América la labor llevada a cabo por el **Tribunal Supremo de los Estados Unidos**<sup>202</sup> con el desarrollo y puesta en marcha del proyecto HERMES para el tratamiento y distribución de las decisiones jurídicas tomadas por dicha institución. Este proyecto fue puesto en marcha en mayo de 1990 con la participación conjunta del citado Tribunal, la Government Printing Office y algunas universidades. En un principio el Tribunal Supremo distribuía sus ficheros electrónicos para su publicación por parte de las instituciones colaboradoras tanto en formato de WordPerfect 5.1 como en texto ASCII, pero desde octubre de 1997 se optó por cambiar el formato de los mismos, realizando la distribución de sus decisiones en el formato Adobe Acrobat PDF y en formato de texto ASCII marcado mediante SGML, forma híbrida entre un marcado descriptivo y un marcado orientado a la presentación. Entre las entidades que distribuían de forma electrónica las decisiones de este Tribunal se encuentra el *Legal Information Institute* de la Cornell Law School de la Universidad de Cornell en el estado de

---

<sup>202</sup> <http://www.supremecourtus.gov/>

Nueva York, que desde 1997 adquiere su propia suscripción al proyecto Hermes comenzando a transformar los ficheros electrónicos recibidos al formato HTML para su publicación en Internet, siendo desde ese momento uno de los sitios más consultados por la comunidad legal norteamericana<sup>203</sup>. En abril de 2000 el Tribunal Supremo de los EE.UU. comenzó a distribuir directamente sus decisiones en Internet a través su propio sitio web.

El caso canadiense no es menos atractivo pues se han dado notables ejemplos del uso del SGML para el tratamiento de la documentación jurídica. Sin duda el caso más importante es el del **Tribunal Supremo de Canadá**<sup>204</sup>, y que por tanto, describiremos brevemente aquí. Como señala D. Poulin, la adopción por parte de este país de la Carta de Derechos y Libertades de los ciudadanos supuso un mayor interés en la población hacia los asuntos legislativos y judiciales<sup>205</sup>. Por ello este Tribunal Supremo decidió poner en marcha un proyecto en colaboración con el *Centre de Recherche en Droit Public* (CRDP) de la Universidad de Montreal para la puesta a disposición del gran público de las colecciones principales de material jurídico existentes en este Tribunal, empleando para ello tecnologías informáticas de gran alcance y distribución, como es el caso de la red Internet. Los trabajos iniciales comenzaron en la primavera de 1994 viéndose la necesidad de establecer un sistema de tratamiento documental que fuese riguroso y preciso. La decisión final fue adoptar el estándar SGML para dicho tratamiento, y se tomó teniendo en cuenta cuatro importantes criterios de partida: la necesidad de simultanear el uso de diferentes tipos de servicios de Internet; el uso de diversos formatos propietarios a parte del tradicional formato ASCII; la implementación de altas capacidades para la búsqueda y recuperación de información y, finalmente, el desarrollo de una secuencia de instrucciones informáticas o *scripts* que

---

<sup>203</sup> Para una mayor información sobre la labor realizada por este Instituto en la difusión electrónica de las decisiones del Tribunal Supremo norteamericano, así como la consulta a sus bases de datos, véase la página web en <http://supct.law.cornell.edu/supct/>

<sup>204</sup> <http://www.scc-csc.gc.ca/>

<sup>205</sup> Daniel Poulin, Alain Lavoie, Guy Huard. "Supreme Court of Canada's cases on the Internet via SGML" [documento HTML]. *E Law – Murdoch University Electronic Journal of Law*, v. 4, n° 3, September 1997. Disponible en <http://www.murdoch.edu.au/elaw/issues/v4n3/poul43.html> (consultado el 8 de enero de 2001).

pudiesen automatizar el procesamiento de los ficheros para su posterior publicación. Se redactó una DTD que expresase adecuadamente las diversas estructuras organizativas de los documentos judiciales existentes en Canadá. Finalmente, desde 1996 estos documentos están a disposición del público a través de la interrogación de sus bases de datos vía Internet, donde los documentos generados en un sistema SGML son convertidos automáticamente al formato HTML cuando el usuario solicita la visualización de los mismos. Reproducimos a continuación la versión simplificada de la DTD desarrollada y expuesta en el artículo de D. Poulin, anteriormente citado, como ejemplo ilustrativo de la mayoría de los desarrollos llevados en todo el mundo para el tratamiento de los documentos jurídicos en los tribunales de justicia<sup>206</sup>:

---

<sup>206</sup> La versión completa de esta DTD se encuentra disponible en <http://www.lexum.umontreal.ca/fr/equipes/technologies/dtd/CSC.dtd>

```

<!DOCTYPE CSC [
<!ELEMENT CSC          - - (SUMMARY, THE.MOTIVES, FINAL.BRIEF)
                        + (PAGE|EMPH) >
<!ELEMENT PAGE        - O EMPTY >
<!ELEMENT SUMMARY     - - (HEADER, BRIEF) >
<!ELEMENT HEADER      - -
(TITLE.C, HEADING, DOCKET.NO, (DATES|BENCH|ORIG) +,
                        ABSTRACT+) >
<!ELEMENT TITLE.C     - - (#PCDATA) >
<!ELEMENT HEADING     - - (CASE, (SEPARATOR, CASE) *) >
<!ELEMENT CASE        - - (PARTY, (SEPARATOR|PARTY) *) >
<!ELEMENT PARTY       - - (PARTY.NAME, STATUS) >
<!ELEMENT PARTY.NAME  - - (#PCDATA) >
<!ELEMENT STATUS      - - (#PCDATA) >
<!ELEMENT SEPARATOR   - - (#PCDATA) >
<!ELEMENT DOCKET.NO   - - (#PCDATA) >
<!ELEMENT DATES       - - (#PCDATA) >
<!ELEMENT BENCH       - - (SEPARATOR|JUDGE) +>
<!ELEMENT (JUDGE|ORIG) - - (#PCDATA) >
<!ELEMENT ABSTRACT    - - (TERM+) >
<!ELEMENT TERM        - - (#PCDATA) >
<!ELEMENT BRIEF       - -
(BRIEF.DECISION, REFERENCES?, FINAL.PAR) >
<!ELEMENT THE.MOTIVES - - (P?, MOTIVE+) >
<!ELEMENT FINAL.BRIEF - - (OUTCOME, PROCURATOR+) >
<!ELEMENT OUTCOME     - - (#PCDATA) >
<!ELEMENT PROCURATOR  - - (#PCDATA) >
]>

```

Figura II.14: Definición de Tipo Documental SGML del Tribunal Supremo de Canadá y el CRDP.

El entorno universitario ha participado también activamente en la aplicación del SGML al tratamiento de la documentación jurídica electrónica, desarrollando proyectos conjuntos con otras instituciones públicas y privadas de este campo de actuación. Son numerosos los proyectos de interés dentro del entorno universitario, casi siempre dirigidos o albergados por facultades o escuelas de Derecho, que se han venido produciendo en estas últimas décadas en diversos países, de los cuales destacaremos los que mayor repercusión y trascendencia han tenido:

El **College of Law Library** de la **Universidad de Cincinnati** en los Estados Unidos es uno de los centros más activos en el estudio e investigación sobre la aplicación de las nuevas tecnologías de la información aplicadas a la documentación de carácter jurídico, y de forma muy especial en el desarrollo de colecciones de material electrónico dentro de las bibliotecas jurídicas. Esta institución creó el prestigioso **Center for Electronic Text in the Law (CETL)**<sup>207</sup> para el estudio y trabajo con materiales digitales desde tres puntos de vista:

- La creación y publicación de recursos digitales para la propia escuela de Derecho.
- La investigación en la búsqueda de los mejores métodos de representación digital de los textos jurídicos.
- La publicación en Internet de documentos jurídicos de especial relevancia.

En el campo específico del tratamiento de los textos legales en formatos electrónicos este centro se ha venido decantando por el empleo y adaptación de la aplicación SGML conocida por el nombre de TEI (*Text Encoding Initiative*) a la documentación jurídica<sup>208</sup>. El CETL ha producido bajo las premisas anteriores dos importantes bases de datos, accesibles a través de Internet. La primera es la base de datos DIANA<sup>209</sup>, única en el mundo por su especialización en materiales jurídicos sobre derechos humanos. La segunda es la denominada *Securities Lawyer's Deskbook*, la cual proporciona acceso electrónico a través de Internet a los textos de la *Securities Act* de 1933 y al *Securities Exchange Act* de 1934. En 1996 esta institución se unió a otras instituciones académicas y de investigación en el campo de

---

<sup>207</sup> El servicio de información en la WWW de este Centro se encuentra en la dirección <http://www.law.uc.edu/CETL/>, aunque es de lamentar la escasa información disponible en este *site*. Gran parte de la información referente a la investigación llevada a cabo por esta institución está disponible, paradójicamente, en otros webs, como se reseñará en las siguientes referencias.

<sup>208</sup> Una de las informaciones más completas sobre la utilización del lenguaje TEI por parte del CETL es facilitada, paradójicamente, por la Universidad de Oxford del Reino Unido, en la dirección de Internet <http://www.hcu.ox.ac.uk/TEI/Applications/index-ce01.html>

<sup>209</sup> Para una mayor información sobre esta base de datos remitimos la consulta de "About DIANA" en <http://elsinore.cis.yale.edu/dianaweb/about.htm>

las bibliotecas jurídicas norteamericanas para formar el *Legal Electronic Text Consortium*<sup>210</sup> con el objetivo principal de la digitalización de materiales de carácter jurídico a través de un desarrollo cooperativo, el mantenimiento a escala internacional de la base de datos DIANA y la elaboración de extensiones de carácter legal de la DTD de SGML para TEI. Entre las instituciones que integran este consorcio destacan las escuelas o facultades de Derecho de universidades tan prestigiosas como la de Columbia, la de Georgetown, la de Nueva York, la de Washington o la de Yale, colaborando asimismo la biblioteca del congreso americano.

El **Centre de Recherche en Droit Public**<sup>211</sup> dependiente de la **Facultad de Derecho** de la **Universidad de Montreal** en Canadá es otro de los centros académicos de mayor prestigio en el terreno de la aplicación de las nuevas tecnologías de la información a la documentación jurídica. El trabajo que este centro ha venido llevando a cabo en la investigación aplicada al tratamiento y difusión de información jurídica y gubernamental es digno de alabanza, pues han sido múltiples y variados los proyectos desarrollados, teniendo como centro la aplicación del lenguaje SGML a dicha información. El trabajo de modelización de la documentación de carácter jurídico de diversas instituciones con las que han venido realizando un sinfín de proyectos (siendo el más destacable el elaborado junto al Tribunal Supremo de Canadá, como se expuso anteriormente) les ha llevado al desarrollo de un gran número de definiciones de tipos documentales de aplicación en las grandes clases de documentos jurídicos: legislación, reglamentación, jurisprudencia y doctrina. Así, uno de los servicios proporcionados por este centro de investigación, el denominado LexUM, ofrece a la comunidad de investigadores y profesionales en este campo un gran número de recursos informativos de todo tipo para el tratamiento de la documentación legislativa y judicial. Para llevar a cabo todo este trabajo, este centro ha venido colaborando estrechamente con otras escuelas y departamentos de la Universidad de Montreal, destacando la estrecha colaboración que se ha venido manteniendo con la escuela de Biblioteconomía y Documentación de la citada universidad.

---

<sup>210</sup> <http://www.law.uc.edu/Diana/letc.html>

<sup>211</sup> <http://www.crdp.umontreal.ca/>

Dentro de los confines del continente europeo destacan sin lugar a dudas dos importantes proyectos de investigación aplicada a la documentación jurídica, llevados a cabo en Finlandia y Suecia. El primero de ellos, el denominado **Raske Project**<sup>212</sup>, no fue encabezado por algún centro procedente del ámbito del derecho sino, más bien, del ámbito de la informática. El **Department of Computer Science and Information Systems** de la **Universidad de Jyväskylä** puso en marcha, junto con el Parlamento de Finlandia y diversos ministerios de este país, este importante proyecto, que ha generado una abundante bibliografía<sup>213</sup>. El proyecto Raske (nombre procedente de las palabras *Rakenteisten AsiakirjaStandardien Kehittäminen*, que puede traducirse como “Desarrollo de Estándares para Documentos Estructurados”), puesto en marcha en 1994, se fundamenta en la investigación y búsqueda de los métodos más eficientes para el tratamiento, gestión y utilización de los documentos de carácter legislativo producidos por el Parlamento y el Gobierno de este país. Sobre la base de un profundo análisis en la tipología documental producida por estas instituciones, se desarrollaron tres DTDs SGML para tres tipos de documentos: proyectos de ley del gobierno, diligencias realizadas por el Parlamento al gobierno, y las réplicas de éste. Estos tipos documentales fueron, además, definidos para generar documentos en diversos idiomas. El primer prototipo, utilizando unas definiciones preliminares de tipo de documento SGML, fue evaluado en 1995<sup>214</sup>. La complejidad de este proyecto ha sido extrema pues siempre ha tenido un planteamiento global de gestión de la información producida por estas instituciones y de mejora de los canales de producción y de comunicación de la misma. Fruto de todo este esfuerzo, el gobierno finlandés ha venido lanzado una serie de desarrollos en la puesta a disposición de material

---

<sup>212</sup> <http://www.cs.jyu.fi/~raske/>

<sup>213</sup> La principal promotora de este proyecto, Airi Salminen, ha publicado un gran número de artículos de diversa índole pero girando muchos de ellos sobre la base de este proyecto. Algunos de estos importantes artículos han sido referenciados con anterioridad en esta tesis.

<sup>214</sup> Jarmo Kiuru, Airi Salminen, Su-Shing Chen. *Digital Archiving of Legislative Documents: A Case Study of Finland* [documento HTML]. Washington D.C.: NSF Workshop on Data Archival and Information Preservation, 1999. Disponible en <http://cecssrv1.cecs.missouri.edu/DA+IPpapers/finland.html> (consultado el 13 de noviembre de 2000).

legislativo electrónico de este país con acceso en línea<sup>215</sup>, entre los que se puede destacar FINLEX<sup>216</sup>. Se trata de una base de datos sobre legislación de este país desarrollada sobre SGML y de acceso gratuito para todos los ciudadanos, puesta en marcha en su primera fase en 1997 y completada en la segunda mitad del año 1999, coincidiendo con la presidencia de la Unión Europea por parte de Finlandia.

El proyecto sueco lleva el nombre de **Corpus Legis Project**<sup>217</sup>, y fue puesto en marcha por el **Swedish Law & Informatics Research Institute** de la **Faculty of Law** con la colaboración del **Department of Computational Linguistics**, ambos pertenecientes a la **Universidad de Estocolmo**. Este proyecto, iniciado en 1994 y finalizado en 1998, tenía como objetivo principal el establecimiento de mecanismos que mejorasen el acceso y recuperación de información jurídica debido, en gran medida, al rápido crecimiento exponencial de información legal así como la internacionalización de la misma necesaria para la armonización europea en esta materia. Al igual que en los anteriores proyectos, la piedra angular de éste giraba en torno a la utilización de sistemas informáticos basados en el estándar SGML por lo que una gran parte del proyecto se basó en el análisis exhaustivo de los corpus textuales de los documentos generados por las distintas instituciones competentes. Tan es así, que el proyecto planteaba las siguientes tres hipótesis de trabajo:

- SGML es una herramienta adecuada para expresar algo más que la estructura de los documentos.
- Es posible diseñar definiciones de tipos de documento (DTDs) para los documentos jurídicos.
- SGML es una herramienta adecuada para el desarrollo de métodos de búsqueda y recuperación de información jurídica.

---

<sup>215</sup> Todos estos desarrollos llevados a cabo por el gobierno de Finlandia se pueden encontrar agrupados en <http://www.sci.fi/~haki/finlaw.html>

<sup>216</sup> <http://finlex.om.fi/>

<sup>217</sup> <http://www.juridicum.su.se/iri/corpus/>

Fruto de todo este trabajo han surgido en estos últimos años diversas publicaciones de gran interés sobre dicha experiencia, destacando la obra de Cecilia Magnusson, una de las investigadoras principales, *Critical Factors in Legal Document Management: A study of standardised markup languages*. Esta obra aborda temas tan importantes como la aplicación de los lenguajes de marcado SGML y XML a los documentos jurídicos, sistemas basados en el hipertexto, sistemas avanzados de recuperación de información, sistemas de gestión electrónica de documentos, así como las técnicas necesarias para el desarrollo de un sistema documental de interés jurídico basado en dichos metalenguajes (análisis documental, diseño de DTDs, marcado de documentos, etc.)<sup>218</sup>.

Por último, y fuera ya del ámbito académico pero dentro del continente europeo, es interesante destacar la importante labor que ha venido desarrollando en el campo de las tecnologías informáticas aplicadas a la documentación jurídica el **Istituto per la Documentazione Giuridica**<sup>219</sup> del **Consiglio Nazionale delle Ricerche** de Italia. Creado en 1968 teniendo el objetivo principal el tratamiento de la documentación de la doctrina jurídica italiana en todas sus facetas, ha venido ampliando su campo de actuación para ser en la actualidad una institución de gran prestigio mundial en el campo de la documentación automatizada de la legislación, de la jurisprudencia y de la doctrina jurídica, así como en el estudio y aplicación de la informática al campo de la actividad jurídica. Entre los numerosos proyectos que ha venido emprendiendo en todos los campos de aplicación de la informática a los textos jurídicos destacamos la intensa labor llevada a cabo en el estudio de instrumentos y métodos adecuados para la construcción de bancos y bases de datos bibliográficos de interés jurídico, y en particular en el estudio de la identificación del dominio documental, el desarrollo de índices de clasificación y tesauros jurídicos especializados, el análisis de programas informáticos de búsqueda y recuperación de

---

<sup>218</sup> Información sobre dicha obra puede encontrarse en la dirección <http://www.juridicum.su.se/iri/corpus/publications.html>, siendo su referencia exacta la siguiente: Cecilia Magnusson Sjöberg. *Critical Factors in Legal Document Management: A study of standardised markup languages*. Stockholm: Jure AB, 1998.

<sup>219</sup> <http://www.idg.fi.cnr.it/>

información, y la aplicación de lenguajes de marcado, como el SGML y el XML, para la construcción de repertorios bibliográficos.

## CONCLUSIONES A LA PARTE

A lo largo de esta parte hemos venido analizando la importancia y las repercusiones de los lenguajes de marcado de textos electrónicos en el desarrollo y avance del tratamiento documental por parte de las instituciones, y de forma muy especial el interés suscitado por el principal lenguaje de marcado de documentos, el SGML (*Standard Generalized Markup Language*). A modo de resumen, y como conclusiones a esta parte se pueden establecer los siguientes puntos de interés:

- El surgimiento de la informática aplicada al control y tratamiento de los textos supuso toda una revolución en el modo de entender los documentos en todas las instituciones. Esta revolución se inició en el entorno de la publicación y edición impresa tradicional en la cual, y con la incorporación de estos mecanismos electrónicos para el tratamiento de los textos que habían de ser publicados, se hacía necesaria la búsqueda de mecanismos que reprodujesen las técnicas tradicionales de trabajo con documentos impresos al entorno de la edición y publicación de los documentos electrónicos. Se hacía pues necesario trasladar el conocimiento acumulado tras siglos de experiencia en la corrección de originales y edición de las pruebas para la impresión a este entorno informatizado de producción documental. La solución vino del empleo de la marca y de lenguajes de marcado del texto electrónico para establecer el modo en el que el texto, y partes concretas del mismo, debía ser procesado por parte de la aplicación informática para su publicación definitiva en un dispositivo de salida específico. Este hecho implicaba asignar un formato o codificación determinada al fichero electrónico que contenía el texto del documento, el cual establecería la forma en la que dicho texto sería presentado. A estos lenguajes orientados exclusivamente al modo en el que el documento debe ser presentado se les conoce por el nombre de lenguaje de marcado de formato.
- Estos lenguajes de marcado de texto poseen serias desventajas desde la óptica del correcto tratamiento y gestión de documentos electrónicos por parte de las

organizaciones: son en su mayoría sistemas propietarios de producción y edición de textos electrónicos, no suministran ningún tipo de información semántica del contenido del documento, no informan sobre los elementos estructurales que definen lógicamente a los documentos, son lenguajes poco flexibles, etc. Frente a este modelo surgen los lenguajes de marcado descriptivo, basados en la premisa de que el contenido del documento debe permanecer separado de su formato de presentación. Las marcas descriptivas utilizadas por estos lenguajes identifican elementos estructurales que componen el documento sin tener en cuenta cómo serán éstos finalmente procesados para su presentación en un determinado medio de salida. Este nuevo modo de entender el documento electrónico reporta innumerables ventajas para cualquier organización: garantiza la independencia informática tanto física como lógica para el tratamiento de los documentos, establece una unidad en el formato del documento electrónico pues lo que se marca son exclusivamente los elementos estructurales que lo componen, permite a las organizaciones establecer una diferenciación de tareas a la hora de producir el documento y el diseño de presentación, garantiza la transportabilidad de los documentos al ser codificado el texto en un sistema universal como es el ASCII, así como otros factores de interés como son la longevidad y perdurabilidad del documento electrónico, la posibilidad de reutilización de partes del documento para generar otros nuevos y la facilidad para la búsqueda y recuperación documental.

- La aplicación del marcado descriptivo a los documentos electrónicos tiene su máxima expresión en el surgimiento de SGML a mediados de la década de los 80. Ideado y concebido inicialmente por Charles F. Goldfarb en los laboratorios de la compañía IBM, es aprobado como estándar internacional ISO en 1986 (norma ISO 8879). SGML es algo más que un lenguaje de marcado descriptivo siendo en realidad un metalenguaje de marcado de documentos electrónicos con una fuerte orientación al desarrollo de sistemas informáticos para la gestión integral de los documentos electrónicos. Como lenguaje de marcado descriptivo que contempla una sintaxis robusta para el establecimiento de marcas en los documentos electrónicos que permitan definir los componentes estructurales de los mismos pero, asimismo, como metalenguaje

contempla una sintaxis abstracta que posibilita a los autores generar sus propios lenguajes de marcado para definir tipos documentales concretos, con todas las ventajas que ello conlleva para cualquier organización que maneje documentos electrónicos. Desde su aparición SGML ha venido siendo completado con otros estándares asociados, los cuales aumentan sus capacidades para, por ejemplo, establecer un potente formato de presentación de los documentos en diferentes medios de salida o crear relaciones hipertextuales de gran complejidad entre documentos SGML.

- La piedra angular del metalenguaje SGML lo constituye la identificación de estructuras lógicas genéricas de los documentos electrónicos y su correspondiente definición a través de este lenguaje. Esta tarea es llevada a través del mecanismo conocido como Definición de Tipo de Documento (DTD). SGML parte de la idea de que los documentos que se manejan en cualquier organización pueden ser agrupados según estructuras de composición lógica similares. Estas estructuras pueden ser definidas mediante un lenguaje determinado, el cual establecerá, entre otras cosas, los nombres para los elementos que componen una determinada estructura lógica genérica, así como la frecuencia de aparición de cada uno de ellos, el orden de aparición de los mismos y el contenido permitido para cada uno de estos elementos. Cada una de estas definiciones de tipos documentales constituye en sí misma un lenguaje de marcado distinto, también denominado *aplicación SGML*. Cada documento SGML contempla, por tanto, una declaración del tipo documental al cual pertenece (estructura lógica genérica) junto con el texto y las marcas empleadas para ese documento en cuestión (estructura lógica específica). Este modo de proceder permite a las organizaciones establecer normas más o menos rígidas para la elaboración de los correspondientes documentos electrónicos fijando unas marcas o otras según el tipo de documento que se está creando. De igual modo, otras ventajas indudables que se derivan de todo ello eson la posibilidad de crear mecanismos que garanticen un intercambio pleno de datos entre organizaciones afines sin pérdidas de información, la reutilización de los documentos, o partes de los mismos, para la generación de otros productos documentales y la capacidad para construir potentes sistemas de búsqueda y recuperación de información electrónica. Esta labor de

definición de tipos documentales a través de SGML no es siempre sencilla pues conlleva un profundo estudio y análisis de la documentación existente dentro de la organización, así como la modelización de la información contenida a través de procesos de transformación de las estructuras físicas de los documentos a sus correspondientes estructuras lógicas genéricas.

- El lenguaje SGML se ha venido aplicando desde su origen a la documentación electrónica generada por organizaciones de todo el mundo y en todo tipo de actividad, no siendo una excepción las instituciones que se inscriben dentro del área de producción documental de carácter jurídico. Aunque con un cierto retraso con respecto a otros sectores productivos, los proyectos y desarrollos que se han venido dando en estas últimas décadas en el campo de la aplicación del SGML al tratamiento de la documentación legislativa y judicial han sido ciertamente notables en todo el mundo, existiendo proyectos de gran envergadura y renombre. En esta Parte de la tesis se han visto desarrollos tan interesantes como los llevados a cabo por ciertas instituciones oficiales de carácter nacional o supranacional dedicadas a la edición y publicación de textos jurídicos emanados de sus correspondientes centros, algunos casos relevantes en los Estados Unidos de América y Canadá para el tratamiento de los documentos electrónicos utilizados en los tribunales de justicia de dichos países, así como algunos proyectos de interés en el continente europeo para el tratamiento de documentos legislativos emanados de sus correspondientes parlamentos y gobiernos. En todos los casos quedan patentes los beneficios que el lenguaje SGML ha reportado, en especial para la racionalización en la producción y gestión de los documentos electrónicos de valor jurídico de estas instituciones elaborando DTDs específicas para cada tipo documental manejado, así como para el desarrollo de sistemas informáticos que posibiliten unas altas capacidades de búsqueda y recuperación documental, siendo numerosas las bases de datos de contenido jurídico confeccionadas a partir de documentos electrónicos marcados con el lenguaje SGML.

- A pesar de las innumerables ventajas aportadas por el lenguaje SGML al tratamiento de la documentación electrónica generada por cualquier tipo de organización, su aplicación no ha sido masiva debido en gran medida a la complejidad que conlleva la implementación de sistemas documentales basados en este lenguaje, así como a los altos costes asociados. Además, desde la aparición del espacio virtual de la red Internet conocido como la *World Wide Web* y el lenguaje que da soporte a los documentos que principalmente circulan por este espacio, el HTML (una aplicación o DTD SGML), su desplazamiento a favor de otras tecnologías ha sido aún mayor debido a los problemas anteriormente citados y a otros relacionados con la falta de adaptación a este entorno informático de comunicación de información.

En la siguiente Parte de la tesis doctoral se analizarán todos estos aspectos, la evolución sufrida por el lenguaje HTML y cómo éste fracasará en el intento de ser un vehículo válido para el tratamiento de los documentos electrónicos complejos que han de circular a través de la red Internet, dando paso a un lenguaje de nueva generación, denominado XML (*Extensible Markup Language*) que promete aunar la potencia del SGML para el tratamiento en la descripción lógica de los documentos electrónicos junto con la sencillez y adaptabilidad al entorno de Internet que posee el HTML. Al igual que en este capítulo se detallarán los proyectos y desarrollos más importantes que se están llevando a cabo en todo el mundo para el tratamiento de la documentación jurídica, y de forma especial dentro de la actividad judicial, en diversos países de nuestro entorno socio-político.



## **PARTE III**

# **XML: LENGUAJE EXTENSIBLE DE MARCADO PARA LA WEB**



## INTRODUCCIÓN A LA PARTE

El surgimiento y desarrollo de los metalenguajes de marcado de texto han significado un auténtico paso de gigante en la búsqueda incesante por parte de los profesionales de la información y la documentación de mecanismos adecuados para el tratamiento y gestión de los documentos electrónicos que se manejan en las organizaciones en las que trabajan, tal como se ha venido exponiendo en la segunda parte de esta tesis doctoral. Pero, sin duda, como ya se apuntó al inicio de dicha parte, será el surgimiento de la red Internet y de forma más concreta el desarrollo y expansión de su servicio más popular, la *World Wide Web* (WWW), la que popularizará los lenguajes de marcado de textos y demostrará la innegable utilidad de éstos tanto en la gestión de la información textual, para la que han nacido y en la que se han desarrollado, como en la gestión tradicional de datos.

El vasto y en buena medida caótico espacio de la WWW ha resultado ser a lo largo de estos años un medio idóneo para el intercambio de ideas y pensamientos entre multitud de comunidades y profesionales de todo el mundo llegando a cotas de implantación mundial jamás pensadas por sus creadores. Todo este inmenso cúmulo de ideas ha venido expresándose a lo largo de todos estos años mediante la publicación de ingentes cantidades de documentos electrónicos marcados mediante las reglas y formas establecidas por el lenguaje de marcas HTML (*HyperText Markup Language*).

Este lenguaje, sencillo de aprender y fácil de aplicar, es una aplicación del metalenguaje SGML al entorno electrónico de la Web. Como tal aplicación, pretendía en sus orígenes conservar las virtudes de su “padre”, es decir, establecer una clara separación entre la estructuración semántica de los contenidos (aunque a un nivel bastante genérico y, por tanto, básico) y la presentación visual de los mismos, pero la evolución que ha venido sufriendo en estos últimos años le ha hecho perder parte de esa pretensión inicial.

En un mundo actual, en el cual la espectacularidad de lo que se nos presenta tiene en ocasiones un mayor peso que el contenido mismo, no resulta extraña esta evolución del servicio Web y, consiguientemente, del mecanismo empleado para crear y difundir esta información electrónica. Además de esto, muchos investigadores y entusiastas de este medio de comunicación y difusión de información han encontrado otras muchas y serias

limitaciones a este lenguaje de marcado, las cuales han puesto en evidencia la necesidad de desarrollar otros lenguajes más potentes y versátiles de aplicación en la futura Web que se pretende construir. Estos nuevos lenguajes vendrán a retomar la filosofía original expuesta años atrás por los creadores de los lenguajes de marcado descriptivo, adaptándose a los nuevos entornos informáticos y de comunicación existentes en la actualidad.

La idea será, como en muchas otras ocasiones en la vida, hacer una revisión histórica de lo sucedido hasta el momento, estudiar y analizar las aportaciones y propuestas que han funcionado mejor y escoger todo aquello que sea satisfactorio para este nuevo entorno de producción, gestión y difusión documental. Si el metalenguaje SGML se reveló como el mejor método para estas tareas documentales, parecía lógico suponer que éste mismo o un metalenguaje derivado de él sería la mejor propuesta posible para crear una nueva Web, donde la estructura y la semántica de los documentos electrónicos prevalecieran sobre otras consideraciones meramente estéticas. El nuevo metalenguaje para la Web toma el nombre de XML (*Extensible Markup Language*). De él se van a derivar un sinnúmero de lenguajes de marcado o aplicaciones específicas para tratar todo tipo de documentos generados en cualquiera de las actividades que desarrolla el ser humano, entre los cuales no serán una excepción los textos jurídicos.

En esta tercera parte de la tesis se tratará de exponer de forma clara y oportuna todas estas ideas anunciadas. Para ello se realizará una revisión histórica desde la aparición y construcción del sistema de tratamiento y gestión de documentos electrónicos que constituye la *World Wide Web* hasta llegar a las principales experiencias y aplicaciones del nuevo metalenguaje XML a los documentos jurisprudenciales. En este largo recorrido, se irán exponiendo algunos de los temas considerados de mayor interés para los propósitos de esta tesis doctoral.

Este camino ha sido largo y complejo y está repleto de infinidad de temas apasionantes, la profundidad y extensión de los aspectos tratados tendrá que ser lo más equilibrada posible, tratando de combinar la brevedad con el siempre necesario rigor científico. Así, a lo largo de esta tercera parte de la tesis doctoral se irá exponiendo el surgimiento y desarrollo de la WWW y de su lenguaje de construcción de documentos electrónicos más característico, el

HTML, analizando la evolución sufrida por éste a lo largo de las últimas décadas y la necesidad de construir algo nuevo que aporte las soluciones requeridas para el nuevo espacio de comunicación electrónica que se desea construir en Internet. La solución encontrada, el metalenguaje XML, uno de los pilares conceptuales de esta tesis doctoral, será tratado en profundidad, analizando aspectos tales como su origen y evolución, la gramática para la construcción de lenguajes específicos de marcado, los beneficios aportados y las actuales limitaciones para su implantación, la complementariedad aportada por otros estándares asociados al mismo y, de forma más profunda, los actuales modelos existentes en este metalenguaje para la definición de estructuras y de contenidos (la semántica de los mismos).

Finalmente, se llegará a la exposición y justificación de la aplicación del metalenguaje XML para el tratamiento de los documentos jurídicos, y de forma más concreta a su utilización en la esfera judicial. En el último capítulo de esta tercera parte se expondrán, por tanto, muchas de las opiniones vertidas sobre este tema por parte de prestigiosos juristas de todo el mundo y se analizarán todas aquellas experiencias y proyectos de aplicación del XML dentro de los tribunales de justicia de diferentes países de nuestro entorno socio-político para el tratamiento de los documentos electrónicos manejados en estas instituciones.



## CAPÍTULO III.1

# LA WORLD WIDE WEB Y EL LENGUAJE

## HTML



### III.1.1. ORIGEN Y EVOLUCIÓN DE LA WWW<sup>1</sup>

Durante la década de los años 70, el área de comunicación y transmisión de datos del CERN (*Centre Européen pour la Recherche Nucleaire*)<sup>2</sup> estaba sumida en un cierto caos tecnológico debido, entre otras causas, a la variedad de técnicas, medios, protocolos de comunicaciones y sistemas operativos diversos existentes dentro de esta institución y sus centros colaboradores<sup>3</sup>. Una de las acciones emprendidas para dar solución a este problema fue la utilización del modelo de red IP de la agencia norteamericana ARPA en el denominado *STELLA Satellite Communication Project*: uno de los canales de este satélite enlazaba diversos nodos de la red CERNET mediante esta tecnología. Este idea se fue extendiendo y en 1984 se creó un proyecto piloto para evaluar el uso de los protocolos TCP/IP en ordenadores del CERN que no utilizaban el sistema operativo Unix (de hecho, el sistema central residía en un *mainframe* IBM-VM y el empleo del sistema VAX/VMS de Digital estaba muy extendido dentro de esta institución)<sup>4</sup>. Aunque el apoyo a esta nueva tecnología no fue inmediato, en 1990 el CERN se convertiría con los años en el centro más

---

<sup>1</sup> Conocer los orígenes y la evolución y difusión de la *World Wide Web* (WWW) y del lenguaje de marcado de hipertexto (HTML) no resulta tarea fácil pues, a pesar de que existe abundante información al respecto, algunos de los datos contenidos en ciertos documentos son contradictorios entre sí, señalado por el mismo Berners-Lee (inventor de la WWW y el HTML) en la página 2 de su libro al comentar textualmente que “La historia de cómo se creó la Web se ha contado en diversos libros y revistas. Muchos relatos que he leído se han distorsionado o son simplemente erróneos”. Por todo ello, y para el correcto desarrollo de este punto, se han tomado como base de trabajo y contrastación de datos los siguientes documentos: diversos documentos electrónicos ubicados en el servidor web del *World Wide Web Consortium* (W3C), en <http://www.w3.org/>, el libro de Tim Berners-Lee. *Tejiendo la Red: El inventor del World Wide Web nos descubre su origen*. Madrid: Siglo XXI de España, 2000, así como el capítulo de 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, Universidad Carlos III de Madrid, 1999, pp. 175-212.

<sup>2</sup> En la actualidad este importante centro europeo de investigación, radicado a las afueras de Ginebra (Suiza), lleva el nombre de *European Laboratory for Particles Physics* aunque el antiguo acrónimo se sigue manteniendo por tradición. Su sitio web se encuentra en <http://www.cern.ch/>

<sup>3</sup> Gregory R. Gromov. *History of Internet and WWW: The Roads and Crossroads of Internet History. Road 2: Europe to USA: Internet at CERN* [documento HTML]. Netvalley.com, 1995, rev, 2000. Disponible en <http://www.netvalley.com/intval2.html> (consultado el 21 de agosto de 2000).

<sup>4</sup> Ben Segal. *A Short History of Internet Protocols at CERN* [documento HTML]. CERN, April 1995. Disponible en <http://wwwinfo.cern.ch/pdp/ns/ben/TCPHIST.html> (consultado el 15 de diciembre de 2000).

grande e importante de Europa en la implantación de las tecnologías de Internet, hecho éste que contribuiría notablemente a la aceptación y expansión de Internet por este continente.

Tim Berners-Lee, empleado del CERN en la división del CN (*Computing & Networks*), había venido trabajando durante la década de los 80 en el desarrollo de un programa de edición electrónica, denominado ENQUIRE (*Enquire Within Upon Everything*), basado en un sistema de relaciones hipertextuales entre documentos producidos por esta organización y que además pudiese trabajar en un entorno informatizado de red Internet, utilizando para ello los protocolos TCP/IP, en los que Berners-Lee estaba especialmente interesado<sup>5</sup>. Fruto de esta experiencia y de sus estudios en el campo de los hipertextos<sup>6</sup>, Berners-Lee lanza en marzo de 1989 una propuesta para la puesta en marcha de un modelo de gestión de información en un entorno informatizado complejo, como era el caso, como se ha comentado con anterioridad, el caso del CERN. El modelo propuesto trataba de crear un sistema informatizado de gestión documental potente donde las pérdidas de información se minimizaran, empleando para ello un sistema distribuido de hipertextos<sup>7</sup>. Paralelamente a esta propuesta, y en ese mismo año, otro investigador del CERN de la división del ECP

---

<sup>5</sup> Dan Connolly. *A Little History of the World Wide Web* [documento HTML]. W3C, rev. December 2, 2000. Disponible en <http://www.w3.org/History.html> (consultado el 2 de febrero de 2001).

<sup>6</sup> No deseamos entrar aquí en la exposición del significado y alcance de los hipertextos pues este tema por sí sólo daría para la realización de una extensa tesis doctoral. Además, otras tesis doctorales que se están desarrollando en el Departamento de Biblioteconomía y Documentación de la Universidad Carlos III de Madrid analizan y exploran con profundidad este tema, desde sus orígenes hasta las actuales propuestas y lenguajes para la construcción de relaciones hipertextuales de gran potencia y versatilidad entre documentos electrónicos. Pero no deseamos dejar de expresar en esta tesis doctoral nuestra gran admiración científica y respeto profesional hacia los considerados padres del hipertexto, Vannevar Bush (con su obra *As We May Think*, publicada a mediados de los 40) y Theodor Holme Nelson (con su obra *Computer Lib*, publicada a mediados de los 60). En cualquier caso, recomendamos para una rápida y sencilla aproximación al concepto e historia de los hipertextos varias obras divulgativas publicadas en nuestro país al respecto del tema; véase José Luis del Río Sadornil. "Los hipertextos, hipermedia, hiperdocumento: una revolución creativa en la información documental". *Documentación de las Ciencias de la Información*, n.º 15, 1992, pp. 83-99, Carles Tomàs i Puig. "Del hipertexto al hipermedia. Una aproximación al desarrollo de las obras abiertas" [documento HTML]. *FORMATS: Revista de Comunicación Audiovisual*, agosto 1997. Disponible en [http://www.iaa.upf.es/formats/formats2/tom\\_e.htm](http://www.iaa.upf.es/formats/formats2/tom_e.htm) (consultado el 21 de agosto de 2000) y, de forma especial, a Mercedes Caridad, Purificación Moscoso. *Los sistemas de hipertexto e hipermedios: Una nueva aplicación en informática documental*. Madrid [etc.]: Fundación Germán Sánchez Ruipérez, 1991.

(*Electronics & Computing for Physics*) , Robert Cailliau, lanza un proyecto de investigación para el tratamiento de la documentación utilizada en esta división basándose, al igual que el anterior, en un sistema de relaciones hipertextuales. El sistema ideado por Berners-Lee cuenta con el apoyo de la dirección de la división del CN (a cargo de Mike Sendall), poniéndose en marcha de forma experimental en un ordenador NeXT (*NeXT cube*, evolución de la primera generación de ordenadores Macintosh); sobre éste Berners-Lee desarrollaría un sistema de edición de documentos hipertextuales y de “navegación” a través de los mismos denominado en, un primer momento, NeXUS (con posterioridad tomaría el mismo nombre que su proyecto).

Las conversaciones informales mantenidas por Berners-Lee y Cailliau en la cafetería del centro supusieron un punto de encuentro entre los desarrollos de ambos investigadores al intentar encontrar un nombre que se ajustase perfectamente al desarrollo que Berners-Lee estaba llevando a cabo; la opción elegida tomó el nombre de **World Wide Web**, frente a otras opciones como las de *Information Mesh*, *Mine of Information* o *Information Mine*<sup>8</sup>. De igual forma, el programa editor/navegador que había desarrollado Berners-Lee en el ordenador NeXT pasaría a denominarse del mismo modo (pero todo junto: *WorldWideWeb*), formando todo ello un único conjunto.

La siguiente imagen muestra una serie de capturas de pantalla de este excelente programa gráfico para la escritura y navegación de documentos hipertextuales:

---

<sup>7</sup> Tim Berners-Lee. *Information Management: A Proposal* [documento HTML]. CERN, March 1989. Disponible en <http://www.w3.org/History/1989/proposal.html> (consultado el 18 de diciembre de 2000).

<sup>8</sup> El propio R. Cailliau comenta que desde un principio desecharon la idea de asignar un nombre procedente de la mitología griega, como era la costumbre hasta ese momento. Berners-Lee propuso el nombre de *World Wide Web* y aunque a Cailliau le pareció un nombre muy acertado, su convencimiento no era total dado que era de difícil pronunciación en francés. Para una mayor información al respecto, véase Robert Cailliau. *A*

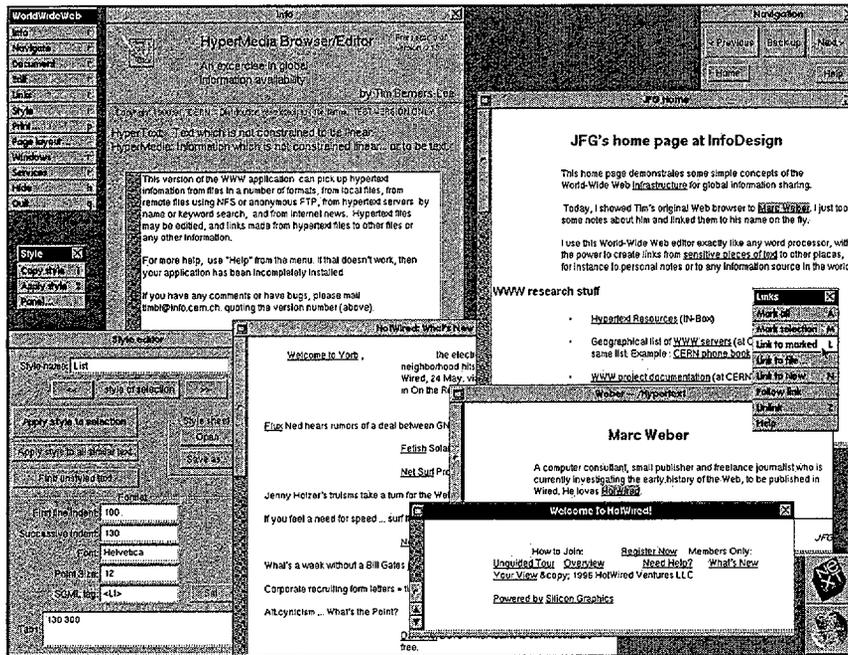


Figura III.1: Editor y navegador gráfico desarrollado por T. Berners-Lee.

Fuente: Tim Berners-Lee. [http://www.w3.org/MarkUp/tims\\_editor](http://www.w3.org/MarkUp/tims_editor)

El 12 de noviembre de 1990 estos dos investigadores lanzan una propuesta conjunta para su debate electrónico entre los miembros de las dos divisiones en las que cada uno de ellos trabajaba, en cuyo título ya se incluía el famoso nombre asignado al modelo experimental: *World Wide Web* (WWW)<sup>9</sup>. En este modelo de entramado informático de redes y ordenadores, cada ordenador cliente se comunica con el ordenador servidor a través del

*Short History of the Web* [documento HTML]. CERN, November 1995. Disponible en <http://www.inria.fr/Actualites/Cailliau-fra.html> (consultado el 15 de enero de 2001).

<sup>9</sup> El documento en cuestión se trataba de un fichero adjunto a un correo electrónico interno a los miembros de las divisiones del CN y del ECP, en el cual se exponían las líneas básicas del modelo de hipertexto propuesto para enlazar información de diversa tipología formando, de este modo, una “tela de araña” (*web*) de nodos residentes en un ordenador servidor entre los cuales el usuario puede moverse a su voluntad desde su ordenador cliente haciendo uso de un programa visualizador (*browser*). Para una más amplia información sobre el tema, véase dicho documento en Tim Berners-Lee, Robert Cailliau. *World Wide Web: Proposal for a HyperText Project* [documento HTML]. CERN, 12 November 1990. Disponible en <http://www.w3.org/Proposal> (consultado el 16 de enero de 2001).

protocolo de comunicación HTTP (*HyperText Transfer Protocol*) para solicitar un determinado documento electrónico, el cual incluye enlaces hipertextuales que ponen su contenido en relación con otros documentos afines. Para conocer la ubicación exacta de los documentos electrónicos en los diferentes servidores web existentes se empleó un sistema de direcciones denominado URI (*Uniform Resource Identifier*)<sup>10</sup>. El lenguaje utilizado para “marcar” los contenidos del fichero electrónico (en formato ASCII) y las relaciones hipertextuales contenidas se denominó HTML (*HyperText Mark-up Language*), aplicación o definición de tipo de documento (DTD) de propósito general derivado del metalenguaje SGML, como se expondrá en el siguiente apartado.

El prototipo desarrollado ofrecía unos resultados asombrosos provocando que a partir del año 1991 se produzca una rápida y creciente difusión del mismo, primero entre los investigadores y directivos de los diversos centros adscritos al CERN (a través de diversas presentaciones y seminarios) y, posteriormente, entre la comunidad científica internacional (a través de la difusión de resultados en diversas listas electrónicas de distribución, presentaciones en congresos internacionales así como la distribución a través de Internet del software empleado en este desarrollo). Sobre el ordenador NeXT de Berners-Lee se registraría el primer servidor Web del CERN cuyo nombre formal era el de *nxoc01.cern.ch* (*NeXT, Online Controls, 1*) registrando el alias de *info.cern.ch*, como nombre alternativo<sup>11</sup>. De igual forma, y a finales de ese año, se pondría en marcha el primer servidor web en los Estados Unidos, ubicado en el *Stanford Linear Accelerator Center* (SLAC) en California.

Debido a que el uso de ordenadores con sistema VMS estaba muy extendido en el CERN el desarrollo tuvo que adaptarse a este sistema, perdiendo gran parte de la potencialidad y vistosidad que le proporcionaba el NeXT (de escasa implantación por su elevado precio).

---

<sup>10</sup> Se puede encontrar abundante información sobre el diseño original de esta primitiva Web en el directorio creado a tal efecto por el *W3 Consortium* en la dirección <http://www.w3.org/DesignIssues/Overview.html>

<sup>11</sup> T. Berners-Lee. *Tejiendo la Red...* *Op. cit.*, p. 26.

Esta adaptación se convertiría en el primer cliente o navegador Web (*web browser*) en modo línea, al cual seguirían con posterioridad otros muchos<sup>12</sup>.

Pero, sin duda, el hecho que contribuyó decisivamente a la rápida expansión y popularidad de la web fue el desarrollo de programas clientes en modo gráfico para la visualización o navegación web instalables en ordenadores con sistemas operativos más populares en todo el mundo<sup>13</sup>. Los primeros que se construyeron fueron para el sistema X-Windows (de gran implantación por aquella época en muchos centros de investigación de todo el mundo) y, posteriormente, para los sistemas Macintosh y Windows. Así, en 1992 se crean en California los visualizadores *Viola* (O'Reilly Associations) y *Midas* (SLAC) para el sistema X-Windows. La demostración realizada en 1993 de estos dos productos en las dependencias del *Software Development Group* de la NCSA (*National Center for Supercomputing Applications*) en Illinois incitó a dos investigadores de dicho centro, Marc Andreessen y Eric Bina, a la construcción de un cliente gráfico para la web de fácil manejo. El resultado fue el desarrollo del *browser Mosaic*, en un primer momento para el sistema operativo Unix y, poco después, para Macintosh y Windows. El hecho de ser un programa de fácil instalación y manejo, robusto y que permitía la visualización de imágenes digitales insertas en los documentos HTML, amén de su distribución gratuita, contribuyó de manera decisiva a la explosión de la tecnología web en los Estados Unidos y, por extensión, en el resto del mundo.

Otros hitos destacables acaecidos durante el año 1993 están en el impulso institucional dado a la WWW por la Comisión Europea, al aprobar a través de su Dirección General

---

<sup>12</sup> Los primeros navegadores para la WWW que se irían desarrollando en los sucesivos años serían, evidentemente, clientes en modo línea. Algunos de los ejemplos más destacados de este tipo de programas lo constituyeron el propio desarrollo del CERN, también denominado *CERN Line Mode Browser*, el navegador *Lynx*, desarrollado por la Universidad de Kansas, o los desarrollos *Emacs W3 Mode*, *Rashty VMS Client* o el *Tom Fine's Perl WWW*. Todos ellos fueron creados, principalmente, para los sistema operativo VMS o Unix, predominantes en ese tiempo. Curiosamente, el navegador desarrollado por Berners-Lee en el CERN no sería el programa más utilizado por la comunidad de usuarios de la WWW durante aquellos años, decantándose éstos de forma mayoritaria por *Lynx*. Para una más amplia información al respecto, véase la obra de John December, Neil Randall. *The World Wide Web Unleashed*. Indianapolis, Indiana: SAMS Publishing, 1994, pp. 137-156.

<sup>13</sup> De igual forma recomendamos la consulta de la obra de December y Randall para una mayor información sobre los desarrollos realizados en este campo. Véase J. December, N. Randall. *Ibid.*, pp.157-280.

XIII el primer proyecto supranacional basado en tecnología web, denominado *Wise*, para la distribución de información electrónica en PYMES. Asimismo, en noviembre de ese año se convoca a los principales investigadores en este campo a la primera reunión internacional WWW en las dependencias del CERN.

En 1994, Andreessen y Bina, seducidos por la oferta de Jim Clark, abandonan la NCSA para fundar *Mosaic Communications*, arrastrando con ellos a un buen número de investigadores que colaboraron en el equipo inicial de desarrollo del navegador Mosaic. Debido a una serie de problemas entre la NCSA y la *Mosaic Communications* por el pago de aranceles del anterior desarrollo, Andreessen y su equipo deciden re-escribir el código fuente y modificar el nombre del producto al de Mozilla, de similar parecido al original pero con grandes mejoras en su funcionamiento<sup>14</sup>. La empresa cambiaría igualmente su nombre por la de *Netscape Communications*.

Con el avance e imparable desarrollo de la WWW, el CERN se encontró desbordado y, consecuentemente, imposibilitado para ser el centro de control y normalización de este complejo entramado mundial que se estaba generando (a finales de 1994 ya había más de 2.500 servidores web en todo el mundo) por lo que tras una entrevista mantenida por Berners-Lee con personal del *Laboratory for Computer Science* del prestigioso MIT (*Massachusetts Institute of Technology*) se vio la necesidad de crear una organización internacional independiente que se hiciera cargo de todo este control. Así, en octubre de 1994 se crea el **World Wide Web Consortium** (W3C) bajo los auspicios del CERN, la agencia norteamericana DARPA y la Comisión Europea, y albergado por tres instituciones de gran prestigio que se repartirían el control y gestión según diversas áreas del planeta: el MIT para el continente americano, el INRIA (*Institut National de Recherche en Informatique et en Automatique*) francés para el continente europeo y la Universidad de Keio en Japón para el continente asiático y las islas del pacífico. Al frente de este consorcio se situará Berners-Lee, el cual abandonaría el CERN para trabajar en el MIT.

---

<sup>14</sup> Gregory R. Gromov. *History of Internet and WWW: The Roads and Crossroads of Internet History. Part 4. Birth of the Web* [documento HTML]. Netvalley.com, 1995, rev, 2000. Disponible en <http://www.netvalley.com/intvalweb.html> (consultado el 21 de agosto de 2000).

Esta organización debería ser más ágil y dinámica en la toma de decisiones que otras organizaciones normalizadoras afines con el fin de promover y expandir rápidamente el uso de la Web por todo el mundo. Citando palabra del propio Berners-Lee al respecto<sup>15</sup>:

*Al igual que el IETF, el W3C desarrollaría especificaciones técnicas abiertas. Contrariamente al IETF, el W3C tendría un personal reducido a tiempo completo para ayudar a diseñar y desarrollar el código cuando fuera necesario. Igual que los consorcios industriales, el W3C representaría el poder y la autoridad de millones de empresarios, investigadores y usuarios. Y al igual que las instituciones miembros de investigación, apoyaría los más recientes avances en tecnología de la información.*

Este Consorcio tiene como misión principal, según figura en su propia documentación, la plena potencialidad de la Web a través del desarrollo de estándares comunes que promuevan su correcta evolución y garanticen su interoperabilidad<sup>16</sup>. Y utilizamos el término “estándares” pues aunque no está legalmente capacitada como organización de ámbito nacional o internacional para el desarrollo de normas oficiales, las “recomendaciones” emitidas por este consorcio tienen a efectos prácticos la misma validez y autoridad que el resto de estándares emitidos por organismos normativos como la ANSI o la ISO. Hasta la publicación de una recomendación del Consorcio Web las opiniones debatidas entre sus miembros pasan por diferentes estadios de trabajo, lo que produce la siguiente tipología documental normativa<sup>17</sup>:

---

<sup>15</sup> T. Berners-Lee. *Op. cit.*, p. 88.

<sup>16</sup> Para una más amplia información sobre la misión, objetivos propuestos y actividades desarrolladas por este Consorcio, véase el documento de Ian Jacobs. *About the World Wide Web (W3C)* [documento HTML]. W3C, March 2000, rev. February 5, 2001. Disponible en <http://www.w3.org/Consortium/> (consultado el 11 de febrero de 2001).

<sup>17</sup> El modo de operar del W3C en relación con el establecimiento de los diferentes procesos y etapas para la adopción de los estándares emitidos, así como la relación de cada uno de los documentos en los que está trabajando o ha trabajado este Consorcio, se encuentra detallado en el documento de Tim Berners-Lee. *W3C Technical Reports and Publications* [documento HTML]. W3C, rev. April 10, 2001. Disponible en <http://www.w3.org/TR/> (consultado el 17 de abril de 2001). En este documento se encuentran asimismo señalados otros documentos de interés elaborados por esta institución relacionados con todo el proceso de establecimiento de normas en el W3C.

1. **Nota** (*Note*): es la propuesta realizada por un organismo o grupo de organizaciones para su debate sin que por ello tenga que haber un compromiso formal para su desarrollo por parte del W3C. Si dicha propuesta es considerada de interés para la comunidad Web se remitirá a un grupo de trabajo del W3C para su discusión formal.
  
2. **Borrador de trabajo** (*Working draft*): es el documento que representa el trabajo que se está llevando a cabo por un determinado grupo auspiciado por el W3C. Estos borradores se publican y difunden entre la comunidad de usuarios de la Web con el fin de fomentar su interés y aportación de ideas por parte de otros grupos involucrados. Suele ser habitual que se produzcan varios borradores de trabajo (*working drafts in development*) hasta llegar a uno final (*working draft in last call*), el cual será revisado por otros grupos de trabajo relacionados con la materia en desarrollo. Asimismo, es posible que algunos de estos borradores de trabajo sean finalmente desestimados o, simplemente, se incorporen dentro de otros documentos como anexos o epílogo del mismo (*working drafts no longer in development*).
  
3. **Recomendación candidata** (*Candidate recommendation*): los borradores de trabajo que han prosperado en su última llamada pasan a tomar el primer nivel para su reconocimiento oficial definitivo. En este caso, la propuesta en curso ha sido revisada y apoyada significativamente por la comunidad de especialistas en la materia abordada por lo que se insta formalmente a otros grupos de trabajo del W3C relacionados con esta materia para que realicen los oportunos comentarios y posibles alteraciones sobre la misma.
  
4. **Recomendación propuesta** (*Proposed recommendation*): con las ideas y propuestas recibidas de los diferentes grupos de trabajo en la fase anterior se produce un consenso final en el seno del grupo de trabajo proponente. De igual modo, este documento ha de ser propuesto formalmente por el Director del Comité Consultivo del W3C para su revisión final. Finalmente, este documento se someterá a la

consideración y votación de las organizaciones miembro del W3C y si es aprobado, pasara al siguiente y último estadio normativo.

5. **Recomendación** (*Recommendation*): es el documento con mayor rango de fuerza del W3C pues representa el consenso final entre todos los organismos, instituciones y personas que integran este consorcio. Este amplio consenso alcanzado es lo que le otorga a este tipo de documentos la consideración *de facto* de estándar internacional en el ámbito de las tecnologías de la información aplicadas a la *World Wide Web*.

El primer encuentro del W3C se produjo el 14 de diciembre de 1994 en Cambridge (Massachussets) en las dependencias del MIT. Desde entonces y hasta nuestros días los encuentros, reuniones y seminarios promovidos y organizados por este Consorcio se han venido sucediendo de forma constante. De igual forma, el interés por la WWW fue rápidamente creciendo por parte de organizaciones públicas y privadas de todo el mundo, en especial compañías de desarrollo de software para Internet que veían en este nuevo espacio electrónico un nuevo y vasto mercado para los negocios; la batalla por su control estaba servida. A partir de esos momentos el desarrollo de la Web será imparable, constituyendo hoy en día el servicio más utilizado dentro de la red Internet por la comunidad internacional de usuarios. La importancia de este servicio es de tal calibre que los intereses políticos, sociales, tecnológicos y económicos que se mueven a su alrededor hacen que su evolución futura sea siempre un reto difícil de prever.

No deseamos finalizar este punto sin ofrecer unas cuantas definiciones más o menos exactas de la *World Wide Web*, tarea ésta no siempre fácil de lograr. Una de las primeras de definiciones establecidas fuera del ámbito del W3C, y que atiende a los orígenes de este servicio, es la proporcionada en el glosario de la NCSA, al definir a éste como “un sistema de información distribuido basado en el hipertexto concebido en el CERN para proporcionar a su comunidad de usuarios un modo sencillo de acceso global a la

información <sup>18</sup>. Pero en la actualidad, y desde una óptica más tecnológica, resulta más adecuada la definición establecida en el glosario del Institute for Telecommunication Science de la NTIA al establecer que se trata de<sup>19</sup>:

*Una red informática virtual de alcance mundial basada en un sistema de información de hipertexto que emplea la Internet como mecanismo de transporte para la visualización de pantallas de ordenador (o páginas web) con información textual, gráfica, imágenes de vídeo e incluso sonora [...] La Web es accesible a través de programas visualizadores (browsers) utilizando el protocolo de transporte de hipertexto (HTTP), un estándar de Internet que especifica cómo una aplicación puede localizar y obtener recursos almacenados en otro ordenador conectado a la Internet. La mayoría de los documentos Web están creados utilizando el lenguaje de marcado de hipertextos (HTML), un sistema de codificación para los documentos WWW. A través de la incorporación del hipermedia (imágenes, sonidos, animaciones, vídeo), la Web se ha convertido en el medio más popular para la publicación de información en Internet.*

Pero más allá de estas definiciones más o menos técnicas, la WWW tiene otro significado y alcance mayor dentro de la sociedad actual, tal vez un tanto alejada de la idea inicial con la que Berners-Lee concibió su lanzamiento dentro del CERN, y que según sus propias palabras sería la siguiente<sup>20</sup>:

*El Web es más una creación social que técnica. Yo lo diseñé por su efecto social —para ayudar a que la gente trabajase junta— y no como un juguete técnico. El objetivo último del Web es apoyar y mejorar nuestra entrelazada existencia en el mundo. Nos agrupamos en familias, asociaciones y empresas. Tenemos confianza en cosas que están a kilómetros y no la tenemos en cosas que están a la vuelta de la esquina. Lo que creemos, aprobamos, aceptamos y de lo que dependemos es representable y, cada vez más, está representado en la Web. Tenemos que asegurar que la sociedad que construimos con el Web es la que pretendemos construir.*

---

<sup>18</sup> *Glossary for NCSA Mosaic and the World Wide Web users* [documento HTML]. Urbana-Champaign, Illinois: The National Center for Supercomputing Applications, rev. June 18, 1997. Disponible en <http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/Glossary/GlossaryDL.html> (consultado el 14 de febrero de 2001).

<sup>19</sup> *Telecom Glossary: World Wide Web* [documento HTML]. Boulder, Colorado: Institute for Telecommunication Science, National Telecommunications and Information Administration, rev. 2000. Disponible en [http://www.its.bldrdoc.gov/projects/t1glossary2000/\\_world\\_wide\\_web.html](http://www.its.bldrdoc.gov/projects/t1glossary2000/_world_wide_web.html) (consultado el 21 de enero de 2001).

### III.1.2. ORIGEN Y EVOLUCIÓN DEL HTML

Antes de comenzar a analizar el origen y la compleja evolución sufrida por el lenguaje de marcado de hipertexto (HTML) resulta conveniente, visto el anterior apartado de este capítulo, ofrecer una definición exacta de este lenguaje de marcado descriptivo. Haciendo referencia a la definición ofrecida por el W3C, responsable actual en el desarrollo y mantenimiento de este estándar, se puede decir que<sup>21</sup>:

*HTML es la lengua franca para la publicación de hipertextos en la World Wide Web. Se trata de un formato no propietario basado en SGML, que puede ser creado y procesado por un amplio rango de herramientas informáticas, desde sencillos editores de texto simple hasta sofisticadas herramientas WYSIWYG. HTML utiliza etiquetas del tipo <b1> y </b1> para estructurar el texto en encabezamientos, párrafos, etc.*

El primer prototipo de sistema de gestión documental desarrollado por T. Berners-Lee en 1989 en los laboratorios del CERN se basaba, como ya se comentó con anterioridad, en la idea de compartir la información generada en las diversas dependencias y centros adscritos a este laboratorio a través de documentos de texto simple que incluyesen relaciones de hipertexto entre los mismos. Ante las dudas suscitadas por este investigador en cuanto a la utilización del mejor sistema para describir los documentos que debían integrarse en este sistema, un colega del laboratorio, Anders Berglund<sup>22</sup>, le recomendó el empleo de una sintaxis basada en SGML. Según palabras de Goldfarb y Prescod, Berners-Lee, tras la consulta y estudio de este estándar internacional y de otros documentos sobre la

---

<sup>20</sup> T. Berners-Lee. *Tejiendo la Red...* Op. cit., p. 115.

<sup>21</sup> Traducción personal de la definición original existente en *HyperText Markup Language. Home Page* [documento HTML]. W3C, rev. February 9, 2001. Disponible en <http://www.w3.org/MarkUp/> (consultado el 14 de febrero de 2001).

<sup>22</sup> La figura de Anders Berglund está, paradójicamente, escasamente reseñada en los documentos que analizan el origen y evolución del lenguaje HTML. Este investigador fue uno de los primeros en la aplicación del metalenguaje SGML a la gestión documental dentro las organizaciones. De hecho, Berglund implantó en 1986 un sistema de gestión basado en SGML en las máquinas IBM del CERN. El apoyo, las ideas y las sugerencias realizadas a Berners-Lee fueron decisivas para la elección del modelo desarrollado.

materia, adaptó un ejemplo de definición de tipo documental (DTD) aparecido en un manual de la compañía IBM, publicado en 1978, para desarrollar su modelo de tipo documental<sup>23</sup>, al cual denominaría **HyperText Markup Language** (HTML).

La idea concebida fue sencilla pero demoledora: basándose en la filosofía de SGML para la construcción de lenguajes de marcado descriptivo, Berners-Lee estableció un pequeño conjunto de etiquetas de marcado que describiesen, por un lado, la estructura básica del documento (por ejemplo, párrafos a través de la etiqueta <P> o la definición de encabezamientos con diferentes niveles a través de las etiquetas <H1> hasta <H6>), que además sirviesen para describir estilos físicos o de presentación de dichos elementos (por ejemplo, el empleo de la cursiva a través de la etiqueta <I> o de texto resaltado en negrilla a través de la etiqueta <B>) y que, por último, fuese capaz de establecer elementos de anclaje (a través de la etiqueta <A>) entre diferentes documentos para la construcción del modelo hipertextual<sup>24</sup>.

El primer borrador de esta sencilla DTD fue implantado en el sistema de edición y visualización *WorldWideWeb*, el cual permitía marcar el texto de los documentos electrónicos de forma rápida y sencilla por parte de cualquier investigador, establecer las oportunas relaciones hipertextuales con otros documentos para, posteriormente, poder acceder cómodamente a la información necesitada y contenida en los documentos relacionados, como se comentó con anterioridad en el apartado primero de este segundo punto del capítulo. Este primer borrador de DTD fue puesto por el CERN a disposición de la comunidad científica internacional para su descarga vía FTP anónimo.

---

<sup>23</sup> Esta aseveración de Goldfarb sobre el origen del lenguaje de marcado HTML, no exenta de cierta polémica, está planteada en la obra de Charles F. Goldfarb, Paul Prescod. *Manual de XML*. Madrid: Prentice Hall Iberia, 1999, p. 13. Ciertamente es difícil pronunciarse sobre este hecho pero, y según nuestro parecer, sí que resulta más creíble que el nacimiento del lenguaje HTML se debiera más bien a un gran conjunto de saberes e influencias tecnológicas (incluida la anotada por Goldfarb) recibidas por Berners-Lee hasta ese momento.

<sup>24</sup> La idea básica de Berners-Lee era la de aunar en un único formato sencillo de manejar la potencialidad de SGML para definir las piezas estructurales del documento, junto con la asignación de un estilo o formato de presentación de los documentos y la inclusión de relaciones hipertextuales entre los mismos (piénsese que en aquella época todavía no se habían desarrollado los estándares asociados a SGML para la definición de estilo, DSSSL, y para el establecimiento de relaciones hipertextuales entre documentos SGML, HyTime).

Desafortunadamente, y según palabras de P. Flynn, este borrador de DTD SGML contenía un gran número de deficiencias que impedían su correcta utilización e implantación<sup>25</sup>. Entre los problemas existentes destacaban, por un lado, la inclusión de algunos errores sintácticos que imposibilitaban que esta DTD fuese analizada por el software SGML existente en el mercado y, por otro, el hecho de que algunos aspectos de esta DTD estaban mal diseñados y, consiguientemente, podrían producir una mala interpretación de los mismos, como era el caso, por ejemplo, de la definición del elemento P (párrafo) como elemento vacío (EMPTY). Realizadas las oportunas correcciones, la nueva versión de esta DTD, no anunciada oficialmente, suscitó tal interés que se produjeron más de cuatro mil descargas del fichero en pocos meses.

En 1991 el CERN sustituyó esta versión por una nueva copia revisada que incluía un conjunto de 25 elementos. Esta versión definitiva del CERN será la que se denomine con posterioridad, de forma no oficial (pues no existía ningún organismo de normalización al cargo de la misma), como **HTML versión 1**<sup>26</sup>.

Dado que el HTML debía de ser servido a través de los protocolos de la red Internet, T. Berners-Lee y R. Cailliau establecieron que el mejor modo para promocionar y asentar todo el entramado HTML creado (el propio lenguaje así como el esquema de direcciones URI y el protocolo de negociación HTTP) debería asentarse en la IETF (*Internet Engineering Task Force*) a través de la creación de un grupo de trabajo específico. Aunque en principio esta institución estaba más interesada por los asuntos técnicos en la normalización de redes de transmisión de datos, como era la Internet, el pujante avance en la instalación de servidores Web en distintas instituciones de gran prestigio por todo el mundo (por ejemplo, el MIT lanzaba por esa época su servidor web con dominio *www.mit.edu*), impulsaría la creación de

---

<sup>25</sup> Peter Flynn. "W[h]ither the Web? The Extension or Replacement of HTML". *Journal of the American Society for Information Science*, v. 48, n° 7, 1997, p. 614.

<sup>26</sup> Margi Mann, Rose Robischon. "HTML Standards – History and Future". *The Serials Librarian*, v. 36, n° 1-2, 1999, p. 53.

diversos grupos de trabajo para la estabilización de los diferentes aspectos relacionados con la WWW, constituyéndose en 1992 el primero de ellos para el tratamiento de los URIs<sup>27</sup>.

El grupo de trabajo de la IETF encargado del estudio del HTML (*HTML Working Group*)<sup>28</sup> comienza su verdadera andadura a comienzos de 1994, no sin ciertos problemas en la tarea encomendada. Por ejemplo, en esos años, diversas instituciones y empresas estaban más interesados en la confección de navegadores para diferentes sistemas operativos (Erwise, Viola y Midas para el sistema X-Windows, Samba, aunque no completado del todo, para Macintosh, Lynx para Unix, y sobre todo el navegador Mosaic de la NCSA para, también, el sistema X-Windows), todos ellos con diferentes versiones del HTML y formatos de estilo o presentación de los contenidos. Se hacía, pues, necesario imponer un cierto orden dentro de esta descontrolada evolución.

A finales de 1993, y en el seno de la primera conferencia internacional WWW desarrollada en el CERN, un investigador de la compañía Hewlett Packard en la delegación de Bristol del Reino Unido, David Raggett, lanza una propuesta para aumentar la potencialidad del lenguaje HTML con la inclusión de otros elementos adicionales, como gráficos, tablas, formularios y la inclusión de fórmulas matemáticas. Este investigador creó dentro de los laboratorios de la citada compañía un navegador web denominado Arena, el cual incorporaba todas estas nuevas posibilidades del HTML. Este importante salto en la potencialidad del HTML para describir nuevos objetos dentro de los documentos hipertextuales fue denominado por su autor como **HTML+**<sup>29</sup>.

---

<sup>27</sup> Por ejemplo, en una de las primeras discusiones técnicas dentro del IETF se estableció el cambio del nombre para el direccionamiento de documentos web, URI, por el de URL (*Uniform Resource Locator*), o localizador uniforme de recursos. T. Berners-Lee. "Tejiendo la Red...". *Op. cit.*, p. 57.

<sup>28</sup> El *HTML Working Group* (HTML-WG) de la IETF ostentó durante los años 1994 a 1996 la responsabilidad en la descripción y desarrollo de este lenguaje de marcado de hipertexto. Su andadura fue corta pero intensa pues supuso una forma de colaboración abierta y menos formal (en comparación con otras instituciones normalizadoras, como la ISO) entre investigadores interesados en este campo. Una mayor información sobre la labor desarrollada por este grupo de trabajo puede ser encontrada en el documento de Dan Connolly. *HTML Working Group of the IETF* [documento HTML]. IETF, rev. January 31, 2000. Disponible en <http://www.w3.org/MarkUp/HTML-WG/Overview.html> (consultado el 18 de diciembre de 2000).

<sup>29</sup> El documento original denominado *HTML+ (Hypertext Markup format)* (disponible en [http://www.w3.org/MarkUp/HTMLPlus/htmlplus\\_1.html](http://www.w3.org/MarkUp/HTMLPlus/htmlplus_1.html)) introducía, además de los nuevos e importantes elementos señalados (imágenes, formularios, tablas, etc.) un mecanismo heredado de SGML para poder

A lo largo de 1994 se trabajó denodadamente dentro del seno del HTML-WG de la IETF en el establecimiento de una normalización entre todas las propuestas que se estaban lanzando y los desarrollos prácticos que se estaban llevando a cabo por esas fechas. Pero no sería hasta noviembre de 1995 cuando ello se conseguiría en cierto grado, fecha en la cual Berners-Lee (que ya trabajaba para el MIT) junto con Connolly lanzaron el denominado RFC 1866<sup>30</sup> dentro esta institución de trabajo y normalización. Tras su rápida aprobación, esta nueva versión del lenguaje de marcado de hipertexto para la web sería conocida como **HTML 2.0**<sup>31</sup>. Esta versión definía ya 50 elementos, añadiendo aquellos aspectos más sobresalientes que establecía el HTML+ para la inclusión de imágenes, tablas y formularios. Aunque la orientación que se le dio a este nuevo desarrollo contaba con ciertas características que lo asentaban mejor en cuanto a la definición de contenidos y la producción de bases de datos, aumentando de este modo las capacidades para la búsqueda y recuperación documental, sus principales problemas vinieron, como señalan Mann y Robischon, en el hecho de que muchos usuarios estaban ya familiarizados con los procesadores de texto WYSIWYG frente a los cuales debería competir este otro sistema de elaboración de documentos electrónicos, con una filosofía diferente y más compleja<sup>32</sup>.

Los elementos definidos en la versión 2.0 del HTML, según P. Flynn, podían ser agrupados dentro de tres categorías: estructurales, descriptivos y visuales<sup>33</sup>. Los elementos estructurales permitían al autor describir la estructura básica del documento, dividido éste

---

“extender” dinámicamente los elementos que puede incluir la DTD del HTML, declarando dichos nuevos elementos como entidades especiales dentro del subconjunto interno de la DTD. Esta particularidad, aunque no prosperaría dentro del lenguaje HTML, sí que será de gran importancia dentro de futuros desarrollos de los lenguajes de marcado para la web, como será el caso del XML. Para una mayor información sobre el HTML+ original y sus diversas modificaciones, véase el documento de David Raggett. *A Review of the HTML+ Document Format* [documento HTML]. Bristol: Hewlett Packard Laboratories, 1994. Disponible en [http://www.w3.org/MarkUp/htmlplus\\_paper/htmlplus.html](http://www.w3.org/MarkUp/htmlplus_paper/htmlplus.html) (consultado el 14 de febrero de 2001).

<sup>30</sup> Tim Berners-Lee, Dan Connolly. *HyperText Markup Language Specification – 2.0* [documento TXT]. IETF, November 1995. Disponible en <http://www.ietf.org/rfc/rfc1886.txt> (consultado el 14 de febrero de 2001).

<sup>31</sup> Toda la información oficial referente a esta nueva versión del lenguaje HTML para la WWW se encuentra en la sección del W3C denominada *HTML 2.0 Materials*, disponible en la dirección <http://www.w3.org/MarkUp/html-spec/index.html>

<sup>32</sup> M. Mann, R. Robischon. *Op. cit.*, p. 54.

<sup>33</sup> P. Flynn. *Op. cit.*, p. 615.

en una cabecera (la cual incluye los elementos para describir bibliográficamente el documento e identificar el material contenido en el texto del documento) y un cuerpo (el cual incluye todo los elementos que serán visualizados por el navegador en la pantalla del ordenador, como cabeceras, listas, párrafos, tablas, formularios, etc.). Los elementos descriptivos permiten mejorar el contenido del documento destacando semánticamente ciertas partes del documento (enlaces hipertextuales, citas, elementos de énfasis, fórmulas matemáticas, abreviaturas y acrónimos, idioma, etc.). Y, por último, otros elementos que afectan directamente a la presentación del texto contenido, mejorando la visualización del mismo, pero sin ningún tipo de connotaciones semánticas (a través del uso de la itálica, la negrita, el subrayado, el tachado, el salto de línea, líneas horizontales, el tamaño y la fuente del texto, etc.).

La irrupción del navegador *Mozilla* de la compañía Netscape a finales de 1994 supuso toda una conmoción en este agitado entorno de desarrollo y normalización del HTML pues desataría una batalla entre diferentes compañías de software por el dominio de este incipiente mercado. Este hecho sería, a la postre, uno de los factores más importantes que desencadenarían la errónea evolución del lenguaje de marcado HTML<sup>34</sup>. Este navegador irrumpió con fuerza, como ya se ha comentado, en un mercado no orientado específicamente a centros de investigación y desarrollo tecnológico. La red Internet y, en especial la Web, empezaba a popularizarse con la creación de servicios web de contenidos netamente comerciales. Este hecho junto con la futura competencia sufrida por otras compañías informáticas desarrolladoras de navegadores web, cambiarán definitivamente el rumbo en la evolución del espacio WWW, produciéndose con posterioridad lo que Siegel denominó la “balcanización” de la Web, o fractura de este espacio en tres mercados distintos en constante pugna, cada uno de ellos con diferentes intereses en el desarrollo y orientación que la Web debería tener: el de la información, el del intercambio de datos y el del ocio y entretenimiento<sup>35</sup>.

---

<sup>34</sup> Brian Kelly. “The Evolution of Web Protocols”. *Journal of Documentation*, v. 55, n° 1, January 1999, p. 72.

<sup>35</sup> Con la popularización de este navegador web este espacio electrónico, que hasta ese momento había sido exclusivo del entorno académico, de investigación y de empresas de comunicación, dicho espacio electrónico

Por otro lado, el gran público deseaba poder crear contenidos para este espacio virtual donde primara, ante todo, la riqueza visual en la presentación de dichos contenidos. Debido a la cierta lentitud con la que se producían los debates en el seno del HTML-WG para la adopción de nuevos elementos, la compañía Netscape decidió empezar a implementar sus propias etiquetas de marcado para ofrecer una mayor capacidad de presentación de contenidos a sus usuarios (por ejemplo, etiquetas como <CENTER> para centrar un determinado elemento o la etiqueta <BLINK> para que un texto pudiera ser destacado a través del parpadeo del mismo). Aunque la compañía Microsoft de Bill Gates desestimó en un primer momento la importancia que estaba tomando la red Internet y su mayor servicio, la WWW, pronto entendió que debería entrar en esta batalla (cada vez eran más los navegadores que se estaban desarrollando para su sistema operativo, el MS-Windows, como era el caso de *Mosaic*, *Cello*, *WinWeb* o el propio *Mozilla* de Netscape). Así, en 1995, tras una serie de duras y agrias batallas con Jim Clark para comprar los derechos de *Mozilla* e integrarlo como navegador de base incluido en el sistema operativo de Microsoft, esta compañía decidió desarrollar su propio navegador, el Internet Explorer, bajo licencia de la versión original del navegador *Mosaic* de la NCSA. Aunque bastante limitado en su funcionalidad y poco competitivo respecto a Mozilla, entró igualmente en esta batalla por el desarrollo de nuevas etiquetas orientadas al enriquecimiento visual del contenido, como fue el caso, por ejemplo, de la etiqueta <MARQUEE> (el texto se desplazaba a lo largo de una marquesina). Netscape reaccionaría con la inclusión de la etiqueta <FRAME> la cual permitía dividir la ventana principal del navegador en diversos marcos independientes.

El HTML-WG trataría de imponer un cierto orden en esta batalla con el lanzamiento por parte de David Raggett, del consorcio W3C, en septiembre de 1995 un *Internet Draft* para el

---

pasaría a convertirse en un medio de masas para el uso y entretenimiento de múltiples grupos de usuarios. El concepto de “balcanización” fue empleado por David Siegel para hacer notar que la fractura producida en este espacio, a partir de ese momento dividido en tres grandes mercados; esto es, el primigenio de la información, junto con el del intercambio de datos entre ordenadores e instituciones y el del ocio y entretenimiento de masas. Del dominio inicial del mercado dedicado a la información y el conocimiento se pasaría a un modelo de espacio web en el que predominan las transacciones económicas y los servicios orientados al ocio y entretenimiento del gran público. Para una mayor información sobre este proceso evolutivo de la Web, véase el documento de David Siegel. “The Balkanization of the Web” [documento

desarrollo de una nueva versión del lenguaje HTML, el **HTML 3.0**<sup>36</sup>, que incluyese los nuevos elementos definidos por los principales navegadores del mercado (en especial las propuestas realizadas por la influyente compañía Netscape para extender este lenguaje con las capacidades que incorporaba su navegador, denominado ya *Netscape Navigator*, en su versión 2.0)<sup>37</sup>. Esta nueva versión no llegó nunca a convertirse en un estándar oficial debido a que fue imposible alcanzar un consenso entre las distintas compañías de informática con intereses en el desarrollo de sus navegadores web. Unido a este hecho, ciertas voces se alzaban contra el modo de proceder de la IETF, donde los debates presenciales y las conversaciones electrónicas del HTML-WG se alargaban interminablemente retrasando el rápido proceso evolutivo que se estaba produciendo en la Web a escala mundial.

Ante este panorama, se decidió que a partir de ese momento la discusión de propuestas y el desarrollo del HTML recayesen en manos exclusivas del W3C, mucho más ágil y dinámica en su forma de funcionamiento y que, además, incorporaba a representantes de todas las compañías de software y hardware con intereses en este asunto. Así, tras unos meses de intensos debates en el seno del W3C se llegó a un cierto consenso, lanzándose en enero de 1997 la siguiente versión de este lenguaje, el **HTML 3.2**<sup>38</sup>. Esta nueva versión, primera verdaderamente oficial del W3C, normalizaba muchas de las propuestas incluidas en la versión 3.0 así como otras nuevas, como el aumento de las capacidades para la producción de tablas complejas, la inclusión de texto flotante alrededor de las imágenes, el

---

HTML]. *Web Techniques*, nº 1, April 1996. Disponible en <http://www.dsiegel.com/balkanization/> (consultado el 23 de enero de 2001).

<sup>36</sup> Dave Raggett. *HyperText Markup Language Specification Version 3.0* [documento HTML]. W3C, September 28, 1995. Disponible en <http://www.w3.org/MarkUp/html3/CoverPage.html> (consultado el 10 de enero de 2001).

<sup>37</sup> Las propuestas lanzadas por esta compañía para esta nueva versión del HTML incluían nuevos elementos orientados principalmente al estilo de presentación del texto, como eran las etiquetas <BIG>, <SMALL>, <SUB> y <SUP>, así como otras capacidades para la creación de mapas de imágenes sensibles a través de la etiqueta <MAP>. Para una mayor información al respecto véase el documento *Extensions to HTML 3.0* [documento HTML]. Netscape, [sin fecha]. Disponible en [http://home.netscape.com/assist/net\\_sites/html\\_extensions\\_3.html](http://home.netscape.com/assist/net_sites/html_extensions_3.html) (consultado el 5 de febrero de 2001).

<sup>38</sup> World Wide Web Consortium. *HTML 3.2 Reference Specification* [documento HTML]. Dave Raggett. W3C, January 14, 1997. Disponible en <http://www.w3.org/TR/REC-html32> (consultado el 5 de febrero de 2001).

desarrollo de mapas de imágenes gestionados por el navegador y, de forma especial, la introducción de ciertos accesorios de programación como los CGI (*Common Gateway Interface*) y, de forma especial, los denominados *scripts* (pequeños programas informáticos desarrollados, en un principio, a través de una versión de Netscape, el *JavaScript*, del lenguaje de programación Java de la empresa Sun Microsystems) los cuales proporcionan un mayor dinamismo y viveza a los documentos HTML.

Este consenso entre los diferentes interlocutores fue más un deseo que una realidad pues ni todos los navegadores existentes en el mercado eran capaces de adaptarse a esta nueva versión (algunos de ellos no reconocían muchas de las capacidades adicionales ahora asignadas) y otros hacían la guerra por su cuenta, caso del Internet Explorer de la compañía Microsoft. Tal es así que esta última compañía desarrolló su versión alternativa a los JavaScripts de Netscape con una versión propia del lenguaje Java, denominada *JScript* ofreciendo, además, a los usuarios del principal lenguaje de programación de esta compañía, el Visual Basic, desarrollar componentes de programación, los *VBScripts*, para este entorno (estos componentes de programación serían denominados con posterioridad *Active X*).

La evolución que estaba tomando el lenguaje de marcado de hipertexto se alejaba cada vez más de la idea original con la que fue concebida, esto es, un lenguaje de marcado descriptivo y sencillo en el que el peso de las marcas recayese en la definición de elementos estructurales y semánticos del texto contenido y no tanto en la presentación del mismo. Durante tiempo atrás se venía hablando de retomar esta idea primitiva y desarrollar, por tanto, un mecanismo por el cual se pudiesen separar estructura y formato de presentación. Influenciados por el entorno de investigación del metalenguaje SGML en este campo a través del desarrollo DSSSL, para la asignación de estilo, como se comentó en el anterior capítulo, el W3C lanzó en diciembre de 1996 la primera recomendación para el uso de hojas de estilo en documentos HTML, conocida como **Cascading Style Shets 1 (CSS1)**<sup>39</sup>. Pero los resultados no fueron los esperados: tan sólo Microsoft a través de, en un principio,

---

<sup>39</sup> World Wide Web Consortium. *Cascading Style Sheets, Level 1. CSS1 Specification* [documento HTML]. Håkon Wium Lie, Bert Bos (eds.). W3C, December 17, 1996, rev. January 11, 1999. Disponible en <http://www.w3.org/TR/REC-CSS1.html> (consultado el 5 de febrero de 2001).

la versión 3.0 del Internet Explorer era capaz de realizar una cierta interpretación de este nuevo lenguaje, siendo omitido por completo por el Netscape Navigator, en su versión 3.0 en esos momentos<sup>40</sup>.

La siguiente versión del lenguaje de marcado de hipertexto, el **HTML 4.0**<sup>41</sup>, comenzó a desarrollarse en julio de 1997 y fue publicada como recomendación en diciembre de 1997. Esta especificación fue revisada en abril de 1998 para corregir algunos errores apreciados en el texto anterior. HTML 4.0 incluyó un gran número de mejoras en diversas áreas, como eran, entre otras, una mayor potencialidad en la definición de tablas y formularios, un mejor soporte para proporcionar contenidos multimedia a través de la Web a discapacitados, establecimiento del estándar ISO 10646 (más conocido bajo el nombre de UNICODE) para dar soporte a textos en cualquier lenguaje humano, el reconocimiento del uso de los *frames* en las páginas web y un mecanismo para dar soporte a las hojas de estilo CSS.

En realidad la versión 4.0 es un conjunto formado por tres DTDs, empleando una u otra según los propósitos del autor; a saber: la DTD estricta (*strict*) es utilizada para indicar al cliente web que dicho documento HTML es estricto con respecto a su marcado y que hace, de igual forma, uso de las hojas de estilo para la presentación de sus contenidos; la DTD transitoria (*transitional*) es la empleada para desarrollar documentos HTML en los que primen las antiguas etiquetas orientadas a la presentación (todas ellas desaconsejadas en esta versión) y, por tanto, para usuarios que utilicen versiones no actualizadas de los navegadores; y la DTD de conjunto de marcos (*frameset*) para la incorporación de marcos o ventanas en el documento HTML.

Para dar un mayor impulso a las capacidades de esta nueva versión del HTML se extendieron las capacidades ofrecidas por el lenguaje de hojas de estilo en cascada,

---

<sup>40</sup> B. Kelly. *Op. cit.*, p. 73.

<sup>41</sup> World Wide Web Consortium. *HTML 4.0 Specification. W3C Recommendation* [documento HTML]. Dave Raggett, Arnaud Le Hors, Ian Jacobs (eds.). W3C, rev. April 24, 1998. Disponible en <http://www.w3.org/TR/1998/REC-html40-19980424/> (consultado el 7 de febrero de 2001).

procediéndose en mayo de 1998 a lanzar la nueva versión de este lenguaje, el **CSS2**<sup>42</sup>, donde destaca, entre otras muchas cosas, la posibilidad de definir diferentes presentaciones del documento según el medio o dispositivo de salida (pantalla del ordenador, impresora o medio sonoro).

Debido a que los lenguajes de programación para este entorno, *JavaScript* y *VBScript*, resultaban ciertamente limitados para establecer un modelo de documento basado en objetos debido principalmente a la ausencia de una descripción normalizada de los distintos componentes que se integraban dentro del documento HTML que fuera más allá de la definición formal de la estructura documental, el W3C desarrolló y lanzó en 1998 la recomendación **DOM** (*Document Object Model*)<sup>43</sup>, confeccionada en dos partes o niveles. DOM proporciona una interfaz independiente de lenguaje de programación para los documentos HTML y hojas de estilo, así como la definición en la forma en que los programas pueden acceder a los datos u objetos contenidos en el documento para modificar el contenido, la estructura y el estilo del mismo. El término **Dynamic HTML** (DHTML) se empleó para describir el uso de un lenguaje de escritura de *scripts* en unión con el DOM con el fin de proporcionar una mayor funcionalidad a las páginas web.

El 22 de diciembre de 1999 el W3C aprobaría la que sería la última y definitiva versión del lenguaje HTML (como tipo documental definido a través de SGML y estandarizada por este consorcio), el denominado **HTML 4.01**<sup>44</sup>, que no es otra cosa que una revisión menor de la anterior en la que se corrigen una serie de errores no detectados en su momento.

---

<sup>42</sup> World Wide Web Consortium. *Cascading Style Sheets, Level 2. CSS2 Specification* [documento HTML]. Bert Bos, Håkon Wium Lie, Chris Lilley, Ian Jacobs (eds.). W3C, May 12, 1998. Disponible en <http://www.w3.org/TR/REC-CSS2/> (consultado el 5 de febrero de 2001).

<sup>43</sup> Véase, respectivamente, World Wide Web Consortium. *Document Object Model (DOM) Level 1 Specification. Versión 1.0* [documento HTML]. Lauren Wood... [et al.] (eds.). W3C, October 1, 1998. Disponible en <http://www.w3.org/TR/REC-DOM-Level-1/> (consultado el 7 de febrero de 2001) y World Wide Web Consortium. *Document Object Model (DOM) Level 2 Core Specification. Versión 1.0* [documento HTML]. Arnaud Le Hors... [et al.] (eds.). W3C, November 13, 2000. Disponible en <http://www.w3.org/TR/DOM-Level-2-Core/> (consultado el 7 de febrero de 2001).

A pesar de todos estos intentos del W3C para orientar de forma correcta el empleo del lenguaje HTML por parte de sus usuarios, la suerte ya estaba echada: las tensiones entre las compañías de desarrollo de software para la web se seguían produciendo constantemente, poniendo éstas el acento ahora en otros aspectos más encaminados al *marketing* y publicidad de sus productos dentro del gran mercado de negocio en que se estaba convirtiendo la WWW más que a proporcionar un verdadero soporte a las actuales versiones del HTML y del lenguaje CSS asociado. Por otro lado, la sencillez primitiva de este lenguaje se había perdido por lo que el proceso de aprendizaje se complicaba y, en muchos casos, se hacía una labor ardua y tediosa. Este hecho ocasionó que muchos usuarios decidieran hacer uso de editores WYSIWYG para simplificar la labor en la creación de páginas web (emulando la sencillez de los modernos procesadores de texto), con la consiguiente pérdida del control sobre el mercado<sup>45</sup>. Muchos de estos editores gráficos incorporaban etiquetas y extensiones propias de la compañía desarrolladora del software encaminados a proporcionar vistosas y llamativas presentaciones en este espacio electrónico, pero no interpretables, lógicamente, por los navegadores de las compañías rivales. La WWW se convirtió, en gran medida, en una nueva “Torre de Babel”.

Sin embargo, para la International Standards Organization (ISO) el lenguaje HTML fue considerado un estándar *de facto* dada su amplia aceptación por la comunidad de usuarios de la WWW, lo que provocaría que esta organización empezase con los procedimientos para su reconocimiento como estándar internacional. Así, la ISO dio rango de norma internacional al lenguaje HTML bajo el convencimiento, como señala acertadamente T. Freter, de que este lenguaje de marcado de documentos para la web persistiría al menos

---

<sup>44</sup> World Wide Web Consortium. *HTML 4.01 Specification. W3C Recommendation* [documento HTML]. Dave Raggett, Arnaud Le Hors, Ian Jacobs (eds.). W3C, December 24, 1999. Disponible en <http://www.w3.org/TR/html4> (consultado el 7 de febrero de 2001).

<sup>45</sup> A este respecto resultan de interés los comentarios realizados por J. Tomás Nogales al señalar que “Aunque los editores de HTML han simplificado notablemente el proceso de composición de páginas web, el conocimiento de las características, posibilidades y limitaciones de este lenguaje [...] se demuestra absolutamente necesario para sacarle el máximo partido a la hora de construir documentos HTML correctos y eficientes”. J. T. Nogales Flores. “La revolución de la World Wide Web”. *Op. Cit.*, p. 183.

otros 25 años<sup>46</sup>. La norma internacional del HTML, basada en la versión 4.01 de la recomendación del W3C, fue aprobada el 15 de mayo del año 2000 bajo el título de ISO/IEC 15445:2000 (E) *Information technology – Document description and processing languages – Hypertext Markup Language (HTML)*<sup>47</sup>. La DTD SGML que define al denominado coloquialmente como **ISO HTML** es más rigurosa que la definida por el W3C, estableciendo ciertas restricciones en el uso y secuencia de elementos HTML (por ejemplo, a un elemento de encabezamiento de nivel 1 -<h1>- nunca le puede seguir un encabezamiento de nivel 3 -<h3>-), debiendo existir entre ambos un encabezamiento de nivel 2 -<h2>-). Con esta acción la ISO ha pretendido tomar las riendas del lenguaje HTML: si el W3C apuesta ahora por otro lenguaje para el espacio electrónico de la Web el desarrollo y evolución del HTML como aplicación del SGML deberá, pues, pasar a manos de la misma organización que vela por el mantenimiento, estabilidad y perdurabilidad de los estándares internacionales derivados del metalenguaje SGML. Y éste parece ser el pacto al cual se ha llegado entre ambas instituciones, la ISO toma las riendas del lenguaje HTML de SGML y sus posibles modificaciones<sup>48</sup> y el W3C se dedicaría de lleno al desarrollo y promoción del metalenguaje XML y sus posibles aplicaciones derivadas. Tal es así que el W3C reformuló el lenguaje HTML 4 como una aplicación de XML, dando lugar en enero del año 2000 a la recomendación **XHTML 1.0** (*Extensible HyperText Markup Language*)<sup>49</sup>, línea por la que el consorcio web va a seguir trabajando en un futuro<sup>50</sup>.

---

<sup>46</sup> Todd Freter. *XML: It's the Future of HTML* [documento HTML]. Sun Microsystems, June 2, 1998. Disponible en <http://www.sun.com/980602/xml/> (consultado el 9 de diciembre de 1999).

<sup>47</sup> International Organization for Standardization. *ISO 15445-2000 (E): Information technology – Document description and processing languages – Hypertext Markup Language (HTML)*. Geneva: ISO, 2000.

<sup>48</sup> Y así está siendo: ya existe una propuesta de corrección sobre el estándar internacional HTML denominada *Draft Technical Corrigendum 1* la cual incide en diversos aspectos técnicos de esta DTD. Para una mayor información véase este documento de trabajo en <http://www.cs.tcd.ie/15445/TC1.html> (consultado el 23 de febrero de 2001).

<sup>49</sup> World Wide Web Consortium. *XHTML 1.0: The Extensible Hypertext Markup Language. A Reformulation of HTML 4 in XML 1.0* [documento HTML]. Steven Pemberton... [et al.]. W3C, January 26, 2000. Disponible en <http://www.w3.org/TR/xhtml1/> (consultado el 25 de septiembre de 2000).

<sup>50</sup> Aunque no nos detendremos en analizar este aspecto, sí resulta interesante entender la dificultad que en aquellos años supuso la búsqueda del criterio más idóneo para la cohabitación de HTML y el recién aparecido

---

metalenguaje XML, aportándose soluciones como la inclusión de datos XML dentro de los documentos HTML ( las denominadas *XML Data Islands*) o la opción finalmente escogida de la reformulación del lenguaje HTML en XML. Para una mayor información sobre este tema, véase Todd Freter. *Op. cit.*, <http://www.sun.com/980602/xml/>

## **CAPÍTULO III.2**

# **XML (EXTENSIBLE MARKUP LANGUAGE): LENGUAJE EXTENSIBLE DE MARCADO**

### III.2.1. LOS ORÍGENES DEL XML<sup>51</sup>

El lenguaje HTML fue muy criticado desde sus comienzos por la comunidad de investigadores y usuarios de SGML, describiéndolo, según comenta el propio Berners-Lee, como un “subproducto inferior” del mismo y proponiendo la completa adopción del SGML para la Web<sup>52</sup>. Al HTML se le achacaba, debido a su extrema sencillez, el hecho de ser un formato incapaz de definir correctamente la estructura y la semántica de muchos de los documentos que se manejaban en cualquier institución, dado que, como sabemos, este lenguaje se trataba de un reducido conjunto cerrado de etiquetas, no de un metalenguaje como era el caso de SGML. La mayor parte de los documentos que se manejaban (y hoy en día con mayor motivo) contenían estructuras complejas que sólo podían ser definidas mediante un lenguaje del tipo de SGML, a través del cual se podían establecer complejos lenguajes de marcado descriptivos específicos (o DTDs) para una correcta definición estructural y semántica de los mismos, como eran los casos de TEI, DocBook o CALS, vistos en el anterior capítulo. Además de esto, la fuerte orientación hacia la presentación de

---

<sup>51</sup> Las diversas traducciones al castellano de las siglas XML que se pueden encontrar en los manuales y fuentes de información electrónica sobre la materia son, en algunos casos, confusas. En primer lugar, el término anglosajón “extensible” no debiera traducirse nunca como “extendible” (Rodríguez Muñoz y Díaz Ortuño) y sí como “extensible” dado que es esta la palabra admitida por la Real Academia Española como adjetivación del verbo “extender” (“extendible” no existe, en todo caso, el término anglosajón podría haberse traducido como “ampliable”, pero la tradición no lo ha impuesto). Pero en gran medida, la principal confusión en la traducción a nuestro idioma del nombre de esta tecnología procede de la ubicación de la palabra “extensible”, situada en algunos casos después de la palabra “lenguaje” (Nogales Flores, Arcineaga – “Lenguaje Extensible de Marcado/Marcas”) o, por el contrario, después de la palabra “marcado” (o “marcas”) (Pitts, Goldfarb, Floyd, Young, Powell – “Lenguaje de Marcado/Marcas Extensible”). Según nuestro parecer situar el adjetivo “extensible” al final puede confundir al usuario pues surge la duda en saber si dicha capacidad de extenderse/ampliarse se está aplicando al lenguaje o, por el contrario, al marcado o las marcas. Si XML es un metalenguaje, esto significa que es el lenguaje el que es extensible (un lenguaje que es capaz de extenderse a sí mismo) y no la marca o el marcado, pues ello es producto o acción de la aplicación de este tipo de lenguajes. Por todo ello, a lo largo de esta tesis doctoral se empleará la traducción de “Lenguaje Extensible de Marcado” por considerarla la más ajustada y precisa para nuestro idioma. Por último, señalar que toda la información oficial relativa al Lenguaje Extensible de Marcado (XML) se encuentra ubicada en el sitio Web del *W3 Consortium*, en la dirección <http://www.w3.org/XML/>

<sup>52</sup> T. Berners-Lee. “Tejiendo la red...”. *Op. cit.*, p. 89.

los documentos que el lenguaje HTML tuvo en sus primeras versiones era, de la misma forma, duramente criticada<sup>53</sup>.

La división e incomprendimientos mutuos entre estos dos mundos eran ciertamente elevadas debido, también y en gran medida, por la forma y filosofía de operar y trabajar de las instituciones y personas que se integraban en dichos espacios: SGML más formal y tradicional frente a la comunidad HTML más dinámica y desenfadada.

Lo que sí que parecía un hecho incuestionable para el entorno de investigadores de SGML era que la WWW era un servicio de Internet que no podían ignorar durante mucho tiempo debido a la rápida difusión y popularidad que estaba alcanzando dentro de la comunidad científica y universitaria. Algunos importantes investigadores del entorno académico universitario que habían venido trabajando en el desarrollo y aplicación de SGML, caso, por ejemplo, de Sperberg-McQueen de la Universidad de Illinois en Chicago (institución en la que se desarrolló la DTD SGML conocida como TEI, comentada en el anterior capítulo), veían necesario un mayor acercamiento de la comunidad SGML al entorno HTML para, de este modo, poder aumentar las capacidades y potencialidad de este importante servicio en Internet que se estaba generando. Implementar todo el SGML en la Web era algo impensable dada la complejidad de este metalenguaje y, por tanto, la dificultad de desarrollar herramientas informáticas capaces de procesar por completo todos los aspectos y variantes del mismo<sup>54</sup>.

Las primeras propuestas de esta integración HTML/SGML pasaban por la necesidad de que tanto el protocolo HTTP como los navegadores Web fueran capaces de dar soporte a los documentos definidos mediante SGML, si no de forma completa sí a través de una

---

<sup>53</sup> En los distintos congresos y asambleas internacionales en torno a SGML que por aquellos años se desarrollaban, normalmente bajo los auspicios de la *Graphic Communications Association* (GCA), eran muchas las voces críticas respecto al empleo del HTML como modelo documental válido para el creciente espacio de la WWW. Por citar un ejemplo, véase lo expuesto al respecto de este tema en el encuentro "SGML Europe'95" por Hasse Haitto, Steve Pepper. "SGML is indispensable: here's living proof" [documento HTML]. En: *SGML Europe Conference* (1995. Gmunden). GCA, 1995. Disponible en <http://www.synex.se/se95.html> (consultado el 8 de enero de 2001).

<sup>54</sup> Para una mayor información sobre la historia e intentos de adaptación del SGML a la WWW, recomendamos la consulta de la obra de Yuri Rubinsky, Murray Maloney. *SGML and the Web: Small Steps Beyond HTML*. Upper Saddle River, NJ: Prentice Hall, 1997.

simplificación de este metalenguaje<sup>55</sup>. Esta solución se veía desde la comunidad Web de difícil implantación pues para la mayoría de los usuarios la simplicidad del HTML les resultaba útil e interesante dado que, y en un gran número de casos, tan sólo interesaban los aspectos estilísticos del documento a la hora de presentarlo en la pantalla del ordenador (se rompía con ello la filosofía de un lenguaje descriptivo). Por otro lado, las principales compañías informáticas que desarrollan software para la Web no estaban dispuestas a invertir grandes sumas de dinero en productos basados en SGML, complejos de producir y, por tanto, de elevado precio final. Si para la mayoría de los usuarios el HTML era más que suficiente para satisfacer sus necesidades de edición electrónica de documentos para este espacio y de consulta de los mismos (el desarrollo de editores HTML y de navegadores web no exigía grandes dificultades técnicas ni elevadas inversiones económicas, existiendo, en muchos casos, productos gratuitos para realizar estas tareas de edición y consulta en la Web)<sup>56</sup>.

Conforme el espacio electrónico de la Web iba creciendo en servidores, número de usuarios y complejidad tecnológica se hacía más palpable que el lenguaje HTML resultaba inapropiado para llevar a cabo todo el tremendo potencial que contenía este servicio en Internet. Además de las limitaciones en cuanto a la definición de estructuras documentales complejas y la descripción semántica de los contenidos, se señalaban otras como las escasas capacidades para ser un formato válido de intercambio de información entre organizaciones en comparación con las DTDs SGML desarrolladas con anterioridad, así

---

<sup>55</sup> C. M. Sperberg-McQueen, Robert F. Goldstein. *HTML to the Max: A Manifesto ofr Adding SGML Intelligence to the World-Wide Web* [documento HTML]. Chicago: Second World Wide Web Conference, October 1994. Disponible en <http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/Autools/sperberg-mcqueen/sperberg.htm> (consultado el 22 de enero de 2001).

<sup>56</sup> E. R. Harold, al describir la génesis del metalenguaje XML, pone especial énfasis en este aspecto al señalar que uno de los factores críticos en la no adopción del SGML completo para la Web fue, sin duda, el elevado precio de los principales productos informáticos existentes en ese momento para la edición y tratamiento de documentos SGML. No sin cierta ironía, este autor señala el precio del producto FrameMaker+SGML de la casa Adobe (valorado en esos momentos en 1.995 dólares) a lo que habría que añadir los costes derivados del resto de piezas que son necesarias para un sistema completo de edición y gestión documental basado en SGML. Para una mayor información, véase Elliot Rusty Harold. *XML: Extensible Markup Language*. Foster City [etc.]: IDG Books Worldwide, 1998, p. 8.

como otros problemas para el almacenamiento, búsqueda y recuperación de información<sup>57</sup>. Las posibilidades que ofrecía el atributo de clase (*class*) dentro del marcado HTML<sup>58</sup> podían aliviar en parte estos problemas pero ni su función original era ésta (se crearon para posibilitar la separación real entre el marcado estructural y la asignación de estilos de presentación a través de CSS) ni solucionaba el problema de fondo en cuanto a la correcta definición estructural y semántica de los documentos electrónicos para la Web.

Otras propuestas lanzadas iban encaminadas hacia modelos mixtos donde la base del sistema documental residente en el ordenador servidor estuviese desarrollado en SGML y, a través de unos procesos de transformación de formatos, dicha información fuese suministrada a los navegadores de los usuarios en formato HTML. Este *modus operandi* (existente en la actualidad para el caso de XML, como se expondrá en este capítulo) no satisfacía a todo el mundo debido a las pérdidas de información (estructural y semántica) que se producían en dicha transformación, amén de ser un método complejo y caro de implantar, sólo al alcance de grandes instituciones<sup>59</sup>.

Ante este panorama que no satisfacía a ninguna de las partes implicadas, el W3C intentaba encontrar una solución acertada. Implementar el SGML completo para la Web, como algunas voces seguían reclamando, se veía, según palabras de Connolly, ciertamente arriesgado: la complejidad de este metalenguaje debido a la gran cantidad de características opcionales lo hacían difícilmente operable en un entorno informatizado no preparado para ello, amén, como ya se ha comentado en varias ocasiones, de los altos costes de producción

---

<sup>57</sup> Stuart L. Weibel. "The World Wide Web and Emerging Internet Resources Discovery Standards for Scholarly Literature". *Library Trends*, v. 43, n° 4, Spring 1995, p. 635.

<sup>58</sup> El atributo genérico *class* posibilita al usuario diferenciar elementos cuyo nombre o identificador genérico sea el mismo, estableciendo clases diferentes para cada uno de ellos. Así, por ejemplo, podemos hacer uso de este atributo dentro del elemento P para indicar que dentro del documento HTML existen párrafos de clase "nota" (<P class="nota">), de clase "especial" (<P class="especial">), "normal" (<P class="normal">), etc., o para diferenciar y destacar cualquier otro aspecto semántico o estilístico de algún otro elemento del HTML.

<sup>59</sup> Lou Burnard. *SGML on the Web: too little too soon, or too much too late?* [documento HTML]. Brusells: SGML User Group Conference, October 1996. Disponible en <http://users.ox.ac.uk/~lou/Belux/cznk1.htm> (consultado el 8 de enero de 2001).

asociados<sup>60</sup>. Desde sus comienzos como consorcio internacional para la industria, el W3C había apostado decididamente por el desarrollo del lenguaje HTML (al fin y al cabo Tim Berners-Lee estaba a la cabeza de dicha institución) pero, de igual forma y al mismo tiempo, se había abierto una puerta al desarrollo paralelo de nuevos lenguajes de marcado que dieran una mejor cobertura al espacio electrónico que se iría desarrollando en un futuro; el metalenguaje SGML siempre estuvo presente entre muchos miembros de este organismo.

De este modo, el W3C se encontraba ante el deseo de retener la esencia inicial de la sencillez y simplicidad del HTML (pues ése había sido uno de los condicionantes para su aceptación por parte del gran público y para pequeños proveedores de información) pero al mismo tiempo era necesario encontrar un modelo de lenguaje de marcado de textos electrónicos que pudiese ser flexible y adaptable a las diversas necesidades de los distintos interlocutores (instituciones académicas y de investigación, empresas comerciales, el gran público, etc.) y que permitiese lo que vino en denominarse la “automatización de la Web”, o la capacidad de automatizar ciertos procesos que se daban en la Web a través de un enriquecimiento semántico de sus documentos que permitiese a ciertos programas informáticos (por ejemplo, los robots de búsqueda de información) un mejor rendimiento<sup>61</sup>. Esta interesante idea sería la que más adelante evolucionase hasta el actual concepto del desarrollo de una *Web Semántica*, como se comentará en posteriores apartados de este capítulo.

Si el problema fundamental del lenguaje HTML era su incapacidad para ser ampliado de forma unilateral por el usuario, y citando las palabras de Bosak y Bray, dos de los padres del nuevo lenguaje del W3C y autores de la idea de una segunda generación de la Web basada en documentos XML, “la solución , en teoría, es muy simple; emplear etiquetas que

---

<sup>60</sup> Dan Connolly, Rohit Khare, Adam Rifkin. *The Evolution of Web Documents: The Ascent of XML* [documento HTML]. Pasadena: California Institute of Technology, January 15, 1998. Disponible en <http://www.cs.caltech.edu/~adam/papers/xml/ascent-of-xml.html> (consultado el 17 de agosto de 2000).

<sup>61</sup> Tim Bray. *Beyond HTML: XML and Automated Web Processing* [documento HTML]. Netscape, 1997. Disponible en [http://developer.netscape.com/viewsource/bray\\_xml.html](http://developer.netscape.com/viewsource/bray_xml.html) (consultado el 23 de junio de 2000).

expresen lo que la información es, y no cómo se debe presentar”<sup>62</sup>; es decir, utilizar un marcado de texto autodescriptivo. Siendo así, sería necesario, pues, construir un modelo de metalenguaje de características similares a SGML, que permitiese a los usuarios crear fácilmente sus propios conjuntos de elementos y que a la vez fuese sencillo de implementar en el espacio electrónico de la Web. Para que todo ello fuera posible de realizar sería, por tanto, requisito imprescindible que el nuevo metalenguaje se generase a través del consenso entre las diferentes instituciones y empresas informáticas con intereses en este campo<sup>63</sup>.

Estas ideas fueron expuestas en verano de 1996 por Jon Bosak, de la empresa Sun Microsystems, ante un grupo de expertos en SGML en la conferencia de Seattle patrocinada por la GCA (*Graphic Communications Association*). Todos los presentes entendieron que la solución debería venir de la mano del metalenguaje SGML, pero de un SGML adaptado a la Web y, por tanto, no un SGML completo, reduciendo aquellas partes más complejas y difíciles de implementar en las aplicaciones web. Entre estos expertos rápidamente se estableció un listado inicial de aspectos no esenciales del metalenguaje SGML para el suministro de información estructurada en la Web<sup>64</sup>, aspecto estos que se expondrán con posterioridad en este capítulo.

De forma paralela al entorno de investigación en SGML, la semilla se plantaba dentro del propio *W3 Consortium*: un grupo de investigadores de diversas instituciones y empresas dirigidos por el mencionado Bosak, expertos en SGML pero que entendían las necesidades y modo de operar en el entorno Web, entre los que se pueden destacar a Dan Connolly por parte del W3C, Tim Bray de las empresas *Textuality* y *Netscape* y C.M. Sperberg-McQuenn

---

<sup>62</sup> Jon Bosak, Tim Bray. “XML and the Second-Generation Web” [documento HTML]. *Scientific American*, May 1999. Disponible en <http://www.sciam.com/1999/0599issue/0599bosak.html> (consultado el 25 de julio de 2000).

<sup>63</sup> Claire Warwick, Elliot Pritchard. “‘Hyped’ text markup language. XML and the future of web markup”. *Aslib Proceedings*, v. 25, n° 5, May 2000, p. 175.

<sup>64</sup> Todd Freter. *XML: Mastering Information on the Web* [documento HTML]. Sun Microsystems, March 10, 1998. Disponible en <http://www.sun.com/980310/xml/> (consultado el 9 de diciembre de 1999).

de la Universidad de Illinois, constituyeron en 1996<sup>65</sup> un grupo de revisión del metalenguaje SGML, denominado originalmente *SGML Editorial Review Board* (SGML ERB) bajo los auspicios del W3C. Según palabras del propio Bosak, la meta de este grupo de trabajo era la de “hacer posible el suministro de estructuras de datos auto-descriptivas con diferentes niveles de profundidad y complejidad para aquellas aplicaciones que requiriesen de dicha estructuración”<sup>66</sup>.

El SGML ERB del W3C comenzó su labor en julio de 1996 con la finalidad de lanzar una nueva especificación que cumpliera una serie de requisitos básicos entre los que destacan el hecho de tener que ser directamente utilizable sobre Internet, ser capaz de dar soporte a una amplia variedad de aplicaciones, ser compatible con el metalenguaje SGML, permitir la generación de textos electrónicos en cualquier lengua humana, facilitar la escritura de programas informáticos que procesen estos documentos los cuales, además, estos documentos deben ser legibles por las personas de forma razonadamente clara y sencilla, y su diseño debe ser formal y conciso, entre otros. En resumen, se buscaba un nuevo lenguaje de marcas que simplificase el metalenguaje SGML y que estuviese especialmente diseñado para las aplicaciones en la Web<sup>67</sup>.

Para llevar a cabo esta misión se constituyó un grupo de trabajo operativo en septiembre de ese mismo año, el *SGML Working Group*, el cual desarrollaría en un par de meses

---

<sup>65</sup> Obsérvese que este grupo de trabajo se constituyó prácticamente un año después de la creación del W3C y, como se verá posteriormente, los frutos de sus trabajos fueron de forma paralela (y en algunos casos anteriores) a las distintas recomendaciones que se irían produciendo en el entorno del lenguaje HTML. Si en su momento no se apostó de forma más decidida por el nuevo lenguaje que se estaba generando hay que achacarlo a las presiones y orientaciones que algunas empresas de la industria informática en la WWW realizaban a este consorcio así como a la rápida y gran aceptación del lenguaje HTML entre el común de los usuarios de la Web.

<sup>66</sup> Jon Bosak. *XML, Java, and the Future of the Web* [documento HTML]. Sun Microsystems, March 10, 1997. Disponible en <http://www.ibiblio.org/pub/sun-info/standards/xml/why/xmlapps.htm> (consultado el 25 de agosto de 2000).

<sup>67</sup> Todos estos objetivos iniciales de diseño se encuentran expuestos en la introducción de la recomendación oficial del nuevo metalenguaje que se desarrollaría: el XML. Estos objetivos de partida son detallados y comentados ampliamente en el manual sobre XML de J. Sturm por lo que recomendamos su consulta para una más amplia información al respecto. Véase Jake Sturm. *Developing XML Solutions*. Redmond, WA: Microsoft Press, 2000, pp. 19-21.

(noviembre del 96) el primer *Working Draft*<sup>68</sup> del nuevo metalenguaje de marcado denominado Extensible Markup Language (XML)<sup>69</sup>. Este borrador de trabajo proporcionaba una completa sintaxis para la producción de documentos XML tanto bien formados como válidos (conceptos éstos que serán analizados en apartados posteriores de este capítulo) estableciendo los mecanismos necesarios para la descripción tanto lógica (a través de la declaración de elementos y atributos) como física (a través de la declaración de entidades y notaciones) de los documentos electrónicos a través de una serie de reglas derivadas del SGML para definir tipos documentales o DTDs (lenguajes específicos) propias del XML.

Este nuevo metalenguaje se dio a conocer a la comunidad científica en la Conferencia sobre SGML convocada por la *Graphic Communication Association* (GCA) celebrada en Boston en ese mismo mes. Junto al *Working Draft* presentado Bosak, se elaboró un breve documento impreso confeccionado a partir de preguntas y respuestas (FAQ) sobre el significado y alcance de XML con la finalidad, según el propio autor, de captar la atención de los expertos asistentes a esta conferencia<sup>70</sup>. Como reconoce DeRose, otro miembro del SGML ERB, las opiniones fueron diversas entre los investigadores de la comunidad SGML asistentes a este congreso pero, en líneas generales, se entendieron los beneficios que se derivaban de esta simplificación del SGML para la Web<sup>71</sup>.

---

<sup>68</sup> World Wide Web Consortium. *Extensible Markup Language (XML). W3C Working Draft* [documento HTML]. Tim Bray, C.M. Sperberg-McQueen (eds.). W3C, November 14, 1996. Disponible en <http://www.w3.org/TR/WD-xml-961114> (consultado el 25 de agosto de 2000).

<sup>69</sup> Como señala el propio Tim Bray en un artículo revisionista elaborado por Denison (otro miembro de SGML ERB) sobre la historia del metalenguaje XML, durante esos meses se barajaron diversos acrónimos para este nuevo desarrollo del W3C; así, en el seno de este grupo se propusieron nombres tales como los de MGML (*Minimal Generalized Markup Language*) o SIMPL (*Simple Internet Markup Protocol*). Para una mayor información al respecto de este y otros aspectos sobre el desarrollo inicial del XML, véase D.C. Denison. *The Road to XML: Adapting SGML to the Web* [documento HTML]. XML.com, October 2, 1997. Disponible en <http://www.xml.com/print/w3j/s1.discussion.html> (consultado el 21 de agosto de 2000).

<sup>70</sup> Este documento fue elaborado en un tono bastante desenfadado y con grandes dosis de ironía hacia el entorno de investigación del SGML por lo que, a buen seguro tuvo que “sorprender” a más de un asistente. El documento en cuestión se encuentra en Jon Bosak. *XML – Questions & Answers* [documento HTML]. International SGML Users’ Group, 1997. Disponible en <http://www.isgmlug.org/n3-1/n3-1-18.htm> (consultado el 1 de septiembre de 1999).

<sup>71</sup> *Ibid.*, <http://www.xml.com/print/w3j/s1.discussion.html>

En enero del siguiente año ya se había construido el primer *parser* o analizador sintáctico para XML. Pero fue, sin duda, en los siguientes meses cuando este nuevo lenguaje de marcado de texto de aplicación al entorno web empezaría a conocerse entre los investigadores de este área de trabajo a través de diversos y seminarios internacionales: en abril de 1997, en el seno de la sexta conferencia internacional sobre la *World Wide Web* en Santa Clara, California se dieron a conocer las nuevas modificaciones realizadas sobre el lenguaje XML. Éste vendría definido en dos partes o *Working Drafts*, el primero de ellos se centraría en la sintaxis de este metalenguaje<sup>72</sup> y el segundo establecería los mecanismos necesarios para la definición de enlaces hipertextuales, tanto simples como complejos, dentro de los documentos XML<sup>73</sup>. El éxito producido fue tan rotundo que en julio de ese año el SGML ERB cambiaría su nombre por el más preciso de XML WG (*XML Working Group*) y el SGML WG cambiaría de igual forma su nombre por el de XML SIG (*XML Special Interest Group*). En este nuevo grupo de trabajo sobre XML (al frente del cual seguiría Jon Bosak) se integrarían representantes de 14 compañías y organizaciones con fuertes intereses en el desarrollo de herramientas informáticas para la nueva web que se estaba diseñando, entre las que cabe destacar a compañías de la talla de Adobe, ArborText, Fuji Xerox, Hewlett-Packard, Microsoft, Netscape, SoftQuad o Sun Microsystems, entre otras<sup>74</sup>.

A lo largo de los siguientes meses del año 1997 se fueron sucediendo de forma continuada diversos borradores de trabajo sobre XML, en un principio dividido en las dos partes antes mencionadas (sintaxis y enlaces hipertextuales)<sup>75</sup> para, con posterioridad,

---

<sup>72</sup> World Wide Web Consortium. *Extensible Markup Language (XML): Part I. Syntax. W3C Working Draft* [documento HTML]. Tim Bray, C.M. Sperberg-McQueen (eds.). W3C, March 31, 1997. Disponible en <http://www.w3.org/TR/WD-xml-lang-970331> (consultado el 25 de agosto de 2000).

<sup>73</sup> World Wide Web Consortium. *Extensible Markup Language (XML): Part II. Linking. W3C Working Draft* [documento HTML]. Tim Bray, Steve DeRose (eds.). W3C, April 6, 1997. Disponible en <http://www.w3.org/TR/WD-xml-link-970406> (consultado el 5 de septiembre de 2000).

<sup>74</sup> D.C. Denison. *Op. cit.*, <http://www.xml.com/print/w3j/s1.discussion.html>

<sup>75</sup> Estos nuevos *Working Drafts* “conjuntos” pueden consultarse en las direcciones <http://www.w3.org/TR/WD-xml-lang-970630> y <http://www.w3.org/TR/WD-xml-link-970731>, respectivamente.

diferenciar ambos aspectos a través de grupos y borradores de trabajo independientes que produjesen sus propias recomendaciones, como se expone en este capítulo.

A la anteriormente citada conferencia le siguieron con posterioridad en los sucesivos meses otras tantas, en especial la organizada por la GCA en septiembre en Sydney bajo el lema ya de “SGML/XML Asia Pacific’97” lugar en el que se anunciaría que para finales de ese año ya se tendría confeccionada la propuesta candidata del XML como recomendación oficial del Consorcio<sup>76</sup>.

Finalmente, y tras votación de los miembros integrantes en la W3C, el 10 de febrero de 1998 XML 1.0<sup>77</sup> fue anunciado como Recomendación oficial del W3C, siendo sus editores Tim Bray de Textuality y Netscape, Jean Paoli de Microsoft y C.M. Sperberg-McQueen de la Universidad de Illinois en Chicago<sup>78</sup>.

A partir de esa fecha, el interés por este nuevo metalenguaje fue creciendo exponencialmente sucediéndose los congresos y encuentros técnicos sobre la materia así como el número de artículos de divulgación publicados. De igual modo, a partir de este año, y de forma rápida y constante, empezaron a desarrollarse una serie de recomendaciones o “estándares acompañantes” del XML que complementan a este metalenguaje en aquellas partes deliberadamente no tratadas, como la asociación de hojas de estilo a los documentos XML o los nombres locales de elementos (*XML namespaces*), por

---

<sup>76</sup> Los últimos borradores de trabajo del metalenguaje XML fueron puestos a disposición del público en agosto y noviembre del año 1997 siendo posible su consulta en las direcciones <http://www.w3.org/TR/WD-xml-970807> y <http://www.w3.org/TR/WD-xml-971117>, respectivamente.

<sup>77</sup> World Wide Web Consortium. *Extensible Markup Language (XML) 1.0. W3C Recommendation* [documento HTML]. Tim Bray, Jean Paoli, C.M. Sperberg-McQueen (eds.). W3C, February 10, 1998. Disponible en <http://www.w3.org/TR/1998/REC-xml-19980210> (consultado el 20 de septiembre de 2000).

<sup>78</sup> No resultó casual que los editores de este nuevo desarrollo del W3C pertenecieran a estas tres instituciones. Se intentaba que las dos compañías de software para la WWW más poderosas en esos momentos, Netscape y Microsoft, se comprometieran al desarrollo conjunto de este nuevo metalenguaje para dicho espacio electrónico. La compañía Microsoft llevaba una cierta ventaja sobre Netscape debido al desarrollo en marzo de 1997 de la primera aplicación interpretable por un navegador Web basada en las primeras ideas sobre XML: el *Channel Definition Format (CDF)*. Incorporado al navegador Internet Explorer 4.0 permitía establecer canales de *Webcasting* a los cuales los usuarios podían suscribirse para obtener información “a la carta”. Según palabras de Bosak y Sperberg-McQueen esta integración de ambas compañías sería beneficiosa, además, debido a que la competencia que se mantendría entre ambas compañías aceleraría la aceptación de la nueva aplicación por parte de otras empresas. Para una mayor información, véase D.C. Denison. *Op. cit.*, <http://www.xml.com/print/w3j/s1.discussion.html>

ejemplo. Asimismo, empezarán a aparecer las primeras aplicaciones o lenguajes específicos de marcado derivados de este metalenguaje, como se expondrá en este capítulo<sup>79</sup>.

En cuanto a la evolución última del metalenguaje XML señalar tan sólo que en los siguientes años se detectaron una serie de errores en la redacción original de la recomendación XML por lo que en agosto de 2000 se presentaría un nuevo *working draft* bajo la dirección de los anteriormente citados editores, en el cual se presentaban a debate dichos errores. Esta revisión daría lugar en octubre de ese año a la denominada segunda, y actualmente vigente, edición (que no versión) de la recomendación XML, a la que se uniría Eve Maler de Sun Microsystems como cuarto editor de la misma<sup>80</sup>.

En referencia a la rivalidad mantenida entre los metalenguajes SGML y XML (y, por extensión, entre las instituciones y personas participantes en ambos desarrollos), ésta se siguió manteniendo durante esos últimos años aunque con una clara ventaja, como se verá posteriormente, a favor del XML. El entorno normativo del SGML, y en concreto Charles F. Goldfarb, nunca se resistió a ver desplazado de la Web a su metalenguaje. De hecho, como ya se señaló en el anterior capítulo, este investigador planteó en 1997 una serie de reformas, conocidas bajo el nombre de *WebSGML*, para introducir este lenguaje como soporte para los documentos manejados en este espacio electrónico. En diciembre de 1998, con la recomendación del XML 1.0 ya aprobada por el W3C, las modificaciones presentadas por Goldfarb fueron completadas y aprobadas por la ISO pasando a formar parte de los anexos del estándar internacional SGML. La honra fue salvada en parte pues junto al anexo K (normativo) que definía esta adaptación del metalenguaje SGML para la Web se introducía un anexo L (de carácter informativo) en el que se desarrollaban los

---

<sup>79</sup> Para una mayor información sobre los importantes avances que se produjeron al respecto a lo largo del año 1998 recomendamos la consulta de Robin Cover. *The Essence and Quintessence of XML. Retrospects and Prospects* [documento HTML]. OASIS, December 31, 1998. Disponible en [http://www.oasis-open.org/html/essence\\_of\\_xml.html](http://www.oasis-open.org/html/essence_of_xml.html) (consultado el 9 de diciembre de 1999).

<sup>80</sup> World Wide Web Consortium. *Extensible Markup Language (XML) 1.0 (Second Edition). W3C Recommendation* [documento HTML]. Tim Bray, Jean Paoli, C.M. Sperberg-McQueen, Eve Maler (eds.). W3C, October 6,

requerimientos necesarios para que el lenguaje XML fuese entendido como un subconjunto de SGML, estableciendo para ello la oportuna declaración SGML que deberían llevar todos los documentos XML<sup>81</sup>.

Esta idea de lucha y supervivencia evolutiva entre formatos de codificación de documentos dentro del espacio de la Web ya fue puesta de manifiesto por Connolly *et al.*, en su famoso, y ya citado con anterioridad, artículo *The Evolution of Web documents: The Ascent of XML*, al señalar que “En el competitivo mercado de las tecnologías de Internet, resulta instructivo considerar el modo en que la Web ha competido y derrotado como especie a otros protocolos [...] En esencia, la aparición de XML sobre el espectro de formatos de datos para la Web acaba con la lucha hacia la consecución de la visión original de la Web establecida por sus creadores”<sup>82</sup>. Esta interesante idea fue ampliada y perfectamente descrita con posterioridad por otros autores, entre los que destacan R. Kahre y A. Rifkin, al aplicar el paradigma de la selección natural de las especies establecida por Darwin dentro de este espacio electrónico de cohabitación de diversos formatos de codificación de documentos electrónicos<sup>83</sup>. Esta evolución en los formatos documentales de la Web hasta llegar en la actualidad a XML ha tenido repercusiones en la sintaxis, en el estilo, en la estructura y en la semántica de los mismos, como así fue establecido por los autores anteriormente citados, como se expondrá igualmente a lo largo de los siguientes apartados de este capítulo.

---

2000. Disponible en <http://www.w3.org/TR/2000/REC-xml-20001006> (consultado el 7 de noviembre de 2000).

<sup>81</sup> Estos dos anexos al estándar internacional SGML pueden ser consultados en cualquiera de los manuales sobre SGML relacionados en la bibliografía, siendo igualmente posible su consulta en formato electrónico en el sitio Web de la compañía de Charles F. Goldfarb, en la dirección <http://www.sgmlsource.com/8879rev/n0029.htm> (consultado el 23 de febrero de 2001).

<sup>82</sup> D. Connolly, R. Khare, A. Rifkin. *Op. cit.*, <http://www.cs.caltech.edu/~adam/papers/xml/ascent-of-xml.html>

<sup>83</sup> Rohit Khare, Adam Rifkin. “The origin of (document) species”. *Computer Networks and ISDN Systems*, v. 30, n° 1-7, 1998, p. 389.

## III.2.2. EL METALENGUAJE XML: DEFINICIÓN, SINTAXIS Y BENEFICIOS E INCONVENIENTES

### III.2.2.1. DEFINICIÓN

XML no es una colección de nuevas ideas sino, como acertadamente es señalado por D. Connolly *et al.*, una correcta selección de ideas de algo ya existente con anterioridad que han venido demostrando su validez y utilidad para la edición y tratamiento de los documentos electrónicos a lo largo de estas últimas décadas<sup>84</sup>. Esto es, XML supone, como exponíamos anteriormente, una simple evolución y no tanto, como algunos autores han pretendido señalar, una auténtica revolución.

Pero, ¿qué es exactamente XML? Para responder a esta cuestión utilizaremos la definición proporcionada por la propia recomendación del W3C, la cual comenta en su resumen inicial, traducido a nuestra lengua, lo siguiente<sup>85</sup>:

*El Lenguaje Extensible de Marcado (XML) es un subconjunto de SGML [...]. Su objetivo es permitir que el SGML genérico pueda ser servido, recibido y procesado en la Web de la misma manera que hoy en día es posible con HTML. XML ha sido diseñado de tal modo que sea fácil de implementar y buscando la interoperabilidad tanto con SGML como con HTML.*

El término clave en esta definición se encuentra en el propio nombre de este desarrollo, esto es, el hecho de ser un lenguaje “extensible” o, dicho de otro modo, ampliable. Este hecho le convierte, por tanto, en un metalenguaje o en un lenguaje para describir múltiples

---

<sup>84</sup> D. Connolly, R. Khare, A. Rifkin. *Op. cit.*, <http://www.cs.caltech.edu/~adam/papers/xml/ascent-of-xml.html>

y diversos lenguajes de marcado de documentos electrónicos orientados cada uno de ellos hacia tipos específicos de contenido<sup>86</sup>. Así pues, no se trata de un lenguaje de marcado predefinido como es el caso de HTML sino que, por el contrario, posibilita a los autores la creación de lenguajes de marcado personalizados para diferentes clases o tipos de documentos, definiendo sus propios nombres de elementos y atributos; esto es, un vocabulario propio. Y esto es así, como es señalado por P. Flynn y otros importantes autores del XML SIG, debido a que este metalenguaje está redactado en SGML, el metalenguaje de marcado por excelencia con rango de norma internacional<sup>87</sup>.

En segundo lugar, se hace referencia explícita al metalenguaje SGML siendo una reducción o simplificación del mismo. Que XML sea similar a SGML implica, por tanto, que XML está orientado a la descripción estructural y semántica de los contenidos de los documentos electrónicos por lo que se beneficiará de todas las ventajas que se desprenden, y ya expuestas en el capítulo anterior, de los lenguajes de marcado descriptivo frente a los orientados al formato de presentación<sup>88</sup>. SGML, como ya se sabe, es capaz de representar estructuras de información ciertamente complejas y está perfectamente adaptado para trabajar con grandes volúmenes de información pero resulta extremadamente difícil de aprender por los usuarios y complejo de implementar en programas informáticos, amén de no haber sido construido pensando en las aplicaciones en línea a través de Internet. Así, XML no es más que una simplificación o subconjunto del metalenguaje SGML (como se verá posteriormente en el apartado correspondiente a la sintaxis XML) especialmente diseñado y optimizado para usarse en un contexto Web, tanto de red informática extensa (Internet) o red local (Intranet). Este hecho conlleva que XML no sólo está influenciado

---

<sup>85</sup> W3C. "Extensible Markup Language (XML) 1.0 (Second Edition)". *Op. cit.*, <http://www.w3.org/TR/2000/REC-xml-20001006>

<sup>86</sup> Natanya Pitts. *XML*. Madrid: Anaya Multimedia, 1999, p. 32.

<sup>87</sup> Peter Flynn... [et al.]. *Frequently Asked Questions about the Extensible Markup Language. The XML FAQ, Version 2.0* [documento HTML]. Dublin: University College Cork, rev. May 30, 2001. Disponible en <http://www.ucc.ie/xml/> (consultado el 18 de junio de 2001).

<sup>88</sup> Elliot Rusty Harold. *XML Bible*. Foster City, CA: IDG Books Worldwide, 1999, p. 4.

por su progenitor SGML sino también por el lenguaje HTML de la Web, aunando, por tanto, potencia y sencillez en un mismo lenguaje<sup>89</sup>.

Podemos afirmar, por tanto, que XML se sitúa a medio camino entre HTML y SGML pero en ningún caso nace para reemplazar a ninguno de estos dos estándares. En el caso de su relación con el lenguaje HTML, como es señalado por Goldfarb, XML fue considerado en su origen por algunos autores de forma errónea como el hermano mayor de HTML<sup>90</sup>. Es verdad que al igual que HTML, XML se diseñó para dar cobertura a los documentos electrónicos que han de circular por la Web y, de igual modo, se basaba en la utilización de marcas para delimitar la posición de los elementos estructurales que definen a los documentos; pero poco más. HTML es, como sabemos, un lenguaje o vocabulario derivado de SGML en el que se establece un conjunto reducido de etiquetas predefinidas y XML es, al igual que SGML, un metalenguaje capaz de definir otros lenguajes específicos especialmente diseñados para operar en la Web, desde la propia adaptación del HTML en el XHTML hasta complejos lenguajes capaces de definir electrónicamente fórmulas matemáticas, representaciones moleculares, notaciones musicales o, en el caso que nos ocupa, textos jurisprudenciales, como se describirá en apartados posteriores.

Como señala T. Bray en una entrevista recogida por la revista *Digital Publishing Technologies*, el lenguaje HTML “es una excelente herramienta para la visualización de documentos hipermedia a través de la red Internet. Sin embargo, XML fue diseñado para proveedores de información electrónica en la Web que desean hacer cosas para las que no fue diseñado HTML”<sup>91</sup>. Este hecho hace que la diferencia más notable de XML con respecto al lenguaje HTML de la Web sea la transición que se realiza desde los documentos hacia las aplicaciones, concepto definido por T. Freter bajo el sugerente término de “weblication”

---

<sup>89</sup> Thomas A. Powell. *HTML: Manual de referencia*. Madrid [etc.]: McGraw-Hill Interamericana de España, 1998, p. 610.

<sup>90</sup> Ch. F. Goldfarb, P. Prescod. *Op. cit.*, p. 17.

<sup>91</sup> “Document Markup”. *Digital Publishing Technologies*, v. 2, n° 4, April 1997, p. 13.

(contracción de los términos *web* y *application*)<sup>92</sup>: XML desplazará a las tradicionales páginas web estáticas creadas mediante HTML orientando éstas hacia aplicaciones web que podrán trabajar en el lado del cliente web reduciendo así el tráfico de datos en Internet. Es más, en este contexto operativo de la Web todos los autores han venido coincidiendo en la profunda transformación que sufrirá este espacio electrónico de acceso y difusión de información. Mediante este nuevo lenguaje la información en este espacio, como señala Goldfarb, podrá ser analizada, extraída, clasificada, diseñada y construida a la medida del usuario para la creación de experiencias Web personalizadas orientadas al usuario final<sup>93</sup>.

Con respecto a SGML sucede algo similar pues XML tampoco lo sustituye, tan sólo lo reduce o simplifica para que sea más fácil su implementación en la Web y, por tanto, más manejable para las herramientas informáticas que trabajan en este entorno. Como acertadamente señala A. de la Rosa, se trata en realidad “de hacer un SGML más comercial, que pueda difundirse más rápidamente. Por esta misma razón, muchas de las posibilidades de SGML se omiten en la especificación XML”<sup>94</sup>.

Hans H. Rath señala que la dificultad de implementar directamente SGML en la Web viene derivada por un gran número de aspectos complejos de este metalenguaje que entran dentro, más bien, del debate académico pero que raramente se utilizan en las aplicaciones del mundo real<sup>95</sup>. Para entender la extrema complejidad y la dificultad de integrarlo dentro de las aplicaciones informáticas es necesario recordar que SGML se creó en los albores del desarrollo de los ordenadores de sobremesa donde se trataba de maximizar las capacidades de los mismos debido a las grandes limitaciones de memoria de procesamiento y de

---

<sup>92</sup> Todd Freter. *Beyond Text and Graphics. XML makes Web pages function like applications* [documento HTML]. Sun Microsystems, April 14, 1998. Disponible en <http://www.sun.com/980414/xml/> (consultado el 6 de noviembre de 2000).

<sup>93</sup> Charles F. Goldfarb. *XML in an Instant: A Non-geeky Introduction* [documento HTML]. OASIS, 1999. Disponible en <http://www.oasis-open.org/html/goldfarb.htm> (consultado el 9 de diciembre de 1999).

<sup>94</sup> Antonio de la Rosa, José A. Senso. “XML como medio de normalización y desarrollo documental”. *Revista Española de Documentación Científica*, v. 22, n° 4, oct./dic. 1999, p. 490.

<sup>95</sup> Hans Holger Rath. *XML: Chance and Challenge for Online Information Providers* [documento HTML]. STEP, 1998. Disponible en [http://www.oasis-open.org/html/xml\\_chance\\_challenge.html](http://www.oasis-open.org/html/xml_chance_challenge.html) (consultado el 9 de diciembre de 1999).

capacidad de almacenamiento en disco duro, proporcionando un complejo conjunto de reglas de minimización y de excepciones de uso<sup>96</sup>. Dado que hoy en día estas limitaciones tecnológicas ya no son un factor tan decisivo, XML se concentra, por tanto, en aquellos aspectos de SGML que proporcionan los mayores beneficios y potencialidad y elimina aquellos otros que son difíciles de utilizar e implementar, como se explicará en este capítulo. Tal es así, que la reducción que se realiza de SGML es ciertamente considerable: si el estándar internacional SGML tiene una extensión entorno a las 500 páginas, la recomendación XML ocupa tan sólo unas 35 páginas<sup>97</sup>.

Pero este hecho no implica necesariamente que XML sea radicalmente más sencillo que SGML pues, aunque es verdad que en este nuevo estándar las reglas sintácticas se simplifican notablemente (lo que constituye el denominado “núcleo XML”), producir documentos XML conllevará en muchos casos la utilización de otras especificaciones asociadas como, por ejemplo, XSL/XSLT para la asignación de estilo o XLink y XPointer para el establecimiento de relaciones hipertextuales entre documentos, que complicarán extremadamente el uso de esta tecnología. Tal es así, que existe el peligro potencial que, como señalan Warwick y Pritchard, esta complejidad real que se le añade a XML a través de estas especificaciones adicionales puede ocasionar que muchos de los actuales usuarios, acostumbrados a la sencillez de HTML, no se interesen por este desarrollo tecnológico y, consiguientemente, su empleo podría ser minoritario<sup>98</sup>.

De igual modo sucede cuando se habla de la madurez de esta tecnología (otro factor determinante para su utilización): el núcleo de XML es un estándar sencillo y estable desde 1998 que se ha venido empleando desde entonces para el desarrollo de múltiples y diversos lenguajes o vocabularios específicos en diversas áreas de aplicación pero, sin embargo, los estándares relacionados están, en algunos casos, aún en fase de consolidación y desarrollo.

---

<sup>96</sup> *HTML, SGML, PDF, XML: What is the difference?* [documento HTML]. ArchiTAG International, 1998. Disponible en <http://www.architag.com/solutions/971206-1.html> (consultado el 9 de diciembre de 1999).

<sup>97</sup> C. M. Sperberg-McQueen. “XML and the Future of Digital Libraries”. *The Journal of Academic Librarianship*, v. 24, n° 4, July 1998, p. 315.

<sup>98</sup> C. Warwick, E. Pritchard. *Op. cit.*, p. 178.

Así, y de forma general, se puede decir que XML es más sencillo que SGML tanto para el usuario que ha de crear documentos electrónicos basados en el marcado del texto como para el software que ha de tratar estos documentos electrónicos pues el analizador o *parser* XML trabajará con un conjunto de reglas significativamente menor. Pero, por el contrario, si las aplicaciones a desarrollar están orientadas a la creación, suministro e intercambio de información en un entorno de red Internet/Intranet la complejidad de ambos metalenguajes es similar dado que en ambos casos es necesario garantizar una absoluta consistencia del formato de datos y documentos<sup>99</sup>.

En resumen, XML toma lo mejor de dos mundos: HTML es “portable” (portátil) y SGML es potente; XML es ambas cosas a la vez<sup>100</sup>. Esta circunstancia convierte a XML, como así lo apunta N. Bradley, en un formato de datos ideal para el almacenamiento de texto estructurado y semiestructurado que ha de ser difundido y publicado en una amplia variedad de medios electrónicos<sup>101</sup>.

---

<sup>99</sup> P.G. Bartlett. “Do you Need XML? A Checklist...” [documento HTML]. En: *SGML/XML Europe 98 Conference* (1998. Paris). GCA, 1998. Disponible en <http://www.infoloom.com/gcaconfs/WEB/paris98/bartlett.HTM> (consultado el 5 de octubre de 2000).

<sup>100</sup> Lorenzo Martínez. “XML, en busca del estándar”. *PC Actual*, n° 104, enero 1999, p. 150.

<sup>101</sup> Neil Bradley. *The XML Companion*. Harlow: Addison-Wesley, 2000, p. 3.

### III.2.2.2. SINTAXIS<sup>102</sup>

Al igual que SGML, el metalenguaje XML presenta su sintaxis a través de reglas de producción escritas con notaciones EBNF (*Extended Backus-Naur Form*). Este sistema permite establecer de forma precisa y sin ambigüedades la sintaxis de XML en contraste con lo que supondría tener que describir dicha sintaxis a través de párrafos explicativos en un idioma determinado. El sistema EBNF no está orientado, por tanto, a su interpretación directa por el ser humano y sí, más bien, como un mecanismo idóneo para que los programas informáticos puedan transformar automáticamente la sintaxis que ha de ser procesada por un ordenador<sup>103</sup>. Así, XML se define en su recomendación oficial a través de 89 reglas de producción EBNF que hacen referencia a todo aquello que está permitido en un determinado lugar de un documento XML<sup>104</sup>. Pero dado que algunas cosas son difíciles de expresar y entender con esta notación gramatical, el W3C optó por incluir en la Recomendación párrafos explicativos que aclarasen el sentido de la mayoría de estas reglas de producción. Sirva como ejemplo las tres siguientes notaciones EBNF, las cuales definen la sintaxis que han de tener las etiquetas de inicio, final y de elemento vacío, respectivamente, en los documentos XML:

```
Stag ::= '<' Name (S Attribute)* S? '>'
ETag ::= '</' Name S? '>'
EmptyElemTag ::= '<' Name (S Attribute)* S? '/>'
```

---

<sup>102</sup> Para entender la sintaxis de XML, y no alargar en exceso este apartado, resulta pues conveniente comparar las semejanzas y diferencias existentes con el metalenguaje SGML, ya estudiado en el capítulo anterior y al cual remitimos al lector para una mayor comprensión de lo aquí expuesto.

<sup>103</sup> Norman Walsh. *A Technical Introduction to XML* [documento HTML]. Arbortext, February 1997. disponible en [http://www.arbortext.com/Think\\_Tank/XML\\_Resources/A\\_Technical\\_Introduction\\_to\\_XML/a\\_technical\\_introduction\\_to\\_xm.html](http://www.arbortext.com/Think_Tank/XML_Resources/A_Technical_Introduction_to_XML/a_technical_introduction_to_xm.html) (consultado el 9 de diciembre de 1999).

<sup>104</sup> Lee Anne Phillips. *Special Edition Using XML*. Indianapolis, Indiana: Que, 2000, p. 28. Esta obra incorpora en su apéndice A, Listado A.1 (pp. 678-691) un listado de todas las reglas de producción EBNF empleadas en XML, así como una serie de comentarios explicativos sobre las mismas. Igualmente interesante y aclaratorio de este modelo de producción gramatical puede resultar el Capítulo 37 (pp. 447-452) del manual de XML de Goldarb y Prescod, citado con anterioridad, en el cual se introduce al lector en el conocimiento y uso de este tipo de reglas de producción.

En el caso de la etiqueta de inicio, ésta deberá comenzar siempre con el ángulo de menor que (el signo < viene entrecomillado representando, por tanto, una cadena constante de valores sensibles al tipo de letra) seguido de un nombre válido (se establecen reglas EBNF para ello) y, en el caso de llevar atributo/s, (el asterisco indica, como ya sabemos que puede no aparecer o aparecer tantas veces como sea necesario) seguido de un espacio en blanco (S) y el atributo o atributos correspondientes (de igual forma se establecen reglas EBNF para los nombres de los mismos y sus posibles valores); finalmente, se puede situar opcionalmente un último espacio en blanco antes del ángulo de cierre (>) de la etiqueta de inicio. La etiqueta final deberá contener la barra oblicua después del ángulo de apertura y antes del nombre del elemento y finalizará con un espacio en blanco opcional y el consiguiente ángulo de cierre. Para la etiqueta que contiene un elemento vacío la regla es similar a la de la etiqueta de inicio (en este caso, única) con excepción de que finaliza con la barra oblicua y el ángulo de cierre.

El concepto general de “documento” para XML es muy similar al establecido en SGML: el documento queda definido tanto por su estructura física como por su estructura lógica. Lógicamente, el documento es dividido en unidades y subunidades denominadas elementos y subelementos (con la posibilidad de incluir atributos dentro de los mismos). Físicamente, el documento puede ser dividido en componentes denominadas entidades (u objetos), los cuales pueden ser almacenados en un mismo fichero o, en muchos casos, en ficheros distintos<sup>105</sup>. XML proporciona una sintaxis formal para la descripción de las relaciones que se establecen entre entidades, elementos y atributos que constituyen el documento XML, proporcionando de este modo un mecanismo para indicar al programa informático cómo ha de reconocer las diferentes partes de las que se compone el documento<sup>106</sup>. Asimismo, SGML incluía dos lenguajes dentro del propio metalenguaje para la construcción de documentos, uno para la propia sintaxis de este metalenguaje y otro para establecer los mecanismos necesarios en la definición de tipos documentales o DTDs, por lo que un

---

<sup>105</sup> N. Bradley. *Op. cit.*, p. 4.

<sup>106</sup> Martin Bryan. “An introduction to the Extensible Markup Language (XML)”. *Bulletin of the American Society for Information Science*, v. 25, n° 1, October/November 1998, p. 11.

documento SGML se consideraba “correcto” cuando se ajustaba a ambos lenguajes. Con XML sucede algo similar pero a diferencia del anterior ahora surgen dos nociones de documento “correcto”: documento bien formado y documentos válidos.

El concepto de **documento bien formado** (*well-formed XML document*) no existía en SGML pues ello implica que no es necesario una DTD para crear y procesar los datos contenidos en el documento XML; éste tan sólo ha de estar creado de forma correcta según las reglas sintácticas básicas de XML<sup>107</sup>. El concepto de “bien formado” procede, como señala M. Johnson del mundo de las matemáticas donde es posible escribir una expresión con la notación sintácticamente correcta pero que no signifique nada (lo mismo ocurre en el lenguaje humano, oral o escrito)<sup>108</sup>. Dicho de otro modo, un documento XML considerado como “bien formado” es simplemente aquel que sigue todas las reglas de redacción y de composición estructural marcadas por la recomendación XML y que, por tanto, es analizado sin producir errores por cualquier *parser* creado a tal fin.

En aras de cumplir con los criterios de simplicidad con los que fue concebido el metalenguaje XML algunas de sus expresiones se regularon de forma más restrictiva que en SGML, reduciendo notablemente sus múltiples posibilidades de expresión. De este precepto surge, por tanto, el concepto de documento “bien formado”, no existente en SGML (todo documento SGML debe inexorablemente ajustarse a una DTD); encontramos aquí la primera gran diferencia entre estos dos metalenguajes.

Las principales reglas que este metalenguaje dicta para considerar a un documento XML como bien formado, y que proporcionan otras cuantas diferencias respecto al modelo de instancia de documento en SGML (las correspondientes a la DTD serán expuestas más adelante), son las siguientes<sup>109</sup>:

---

<sup>107</sup> P.G. Bartlett. *Op. cit.* <http://www.infoloom.com/gcaconfs/WEB/paris98/bartlett.HTM>

<sup>108</sup> Mark Johnson. *XML for the absolute beginner* [documento HTML]. JavaWorld, April 1999. Disponible en [http://www.javaworld.com/javaworld/jw-04-1999/jw-04-xml\\_p.html](http://www.javaworld.com/javaworld/jw-04-1999/jw-04-xml_p.html) (consultado el 14 de febrero de 2001).

<sup>109</sup> Resulta del todo imposible detallar de forma exhaustiva en este apartado todas y cada una de las diferencias entre los metalenguajes SGML y XML y, por tanto, describir las reglas sintácticas descritas en la recomendación XML del W3C. Por dicho motivo sólo se destacarán aquellas consideradas como principales y

- Todos los documentos XML llevan implícita una declaración SGML que es fija para todas las instancias de documento dado que muchas de las características opcionales que podían llevar los documentos SGML (como es el caso de los parámetros de minimización del etiquetado *-DATATAG, OMITAG-* o de los delimitadores de referencia breve *-SHORTREF, RANK, LINK, etc.-*, por citar algunos de ellos)<sup>110</sup> toman ahora un valor fijo. Dado que dicha declaración es común para todos los documentos XML ésta puede reducirse a una simple declaración de marca obligatoria, cuya expresión mínima vendría, por tanto, dada por la instrucción de procesamiento siguiente:

```
<?xml version="1.0"?>
```

- En XML, una instrucción de procesamiento (IP) se emplea, en palabras de Goldfarb y Prescod, cuando resulta necesario indicar una ejecución de procesamiento a un programa informático sin tener que hacer referencia a la DTD e, igualmente, sin tener que cambiar la forma en que otros programas informáticos procesan el documento en cuestión<sup>111</sup>. Una IP se construye incluyendo dicha instrucción entre las marcas inicial “<?” y la final “?>”. Un caso muy habitual en este tipo de instrucciones es la derivada del procesamiento de una hoja de estilo asociada a un documento XML, como se expone en el siguiente ejemplo:

```
<?xml-stylesheet type="text/css" href="estilo1.css"?>
```

---

más necesarias, remitiendo al lector además de a la propia recomendación XML (siempre árida de leer y difícil de entender) o a cualquiera de los excelentes manuales existentes en el mercado y referenciados en esta tesis doctoral, al documento elaborado por J. Clark bajo los auspicios del W3C en el que se destacan dichas diferencias existentes. Para ello, véase James Clark. *Comparison of SGML and XML. W3C Note* [documento HTML]. W3C, December 15, 1997. Disponible en <http://www.w3.org/TR/NOTE-sgml-xml-971215> (consultado 7 de noviembre de 2000).

<sup>110</sup> Recuérdese lo comentado en la nota 107 al respecto del anexo L de la norma internacional SGML, la cual establecía de forma informativa una declaración SGML para todos los documentos XML. De igual modo, la especificación XML definía en sus primeros borradores una declaración SGML para los documentos XML. Esta declaración fue eliminada del texto de la Recomendación quedando inscrita dentro del documento de J. Clark, citado en la anterior nota. En dicho documento se establece de igual modo una declaración SGML para XML con algunas pequeñas diferencias con respecto a lo marcado por la ISO en el año L de la norma internacional 8879:1986.

<sup>111</sup> Ch. F. Goldfarb, P. Prescod. *Op. cit.*, p. 438.

- Dado el carácter universal de XML, todo documento XML es un documento de texto formado por caracteres procedentes del conjunto definido por el estándar Unicode<sup>112</sup> (este mismo estándar es publicado por la ISO como la norma internacional ISO/IEC 10646, con ligeras variaciones). XML soporta por defecto el juego de caracteres conocido como ISO-8859-1, correspondientes al juego de caracteres empleados en Europa occidental (ISO-8859-1 o también denominado ISO-Latin-1) pero debe ser capaz de dar soporte a los subconjuntos de caracteres UTF (*Unicode Transformation Format*) de Unicode, tanto de 8 bits (UTF-8) como de 16 bits (UTF-16). Este hecho no se ha logrado verdaderamente hasta las modificaciones realizadas en la segunda edición de la Recomendación XML del W3C<sup>113</sup>. El juego de caracteres Unicode empleado en el documento XML debe hacerse constar (siempre que no sea el definido por defecto) en la declaración XML correspondiente, como se muestra en el siguiente ejemplo:

```
<?xml version="1.0" encoding="UTF-8" ?>
```

- A semejanza de SGML, en todo documento XML debe existir un único elemento de documento (elemento raíz) en el que todos los demás elementos y contenidos se encuentren anidados; por ejemplo:

```
<documento>  
<cabecera>.....</cabecera>  
<cuerpo>.....</cuerpo>  
</documento>
```

- Todos los elementos, atributos y entidades que se vayan a emplear en un documento XML han de seguir una serie de reglas sintácticas; a saber:

---

<sup>112</sup> Para mayor información sobre este estándar, actualmente en la versión 3.2, véase la información suministrada en el sitio Web del consorcio internacional que actualmente lo mantiene en <http://www.unicode.org/>

<sup>113</sup> De hecho, en la edición de 1998 de la Recomendación XML no se describe a este metalenguaje como una aplicación del estándar Unicode por lo que existían, como señala al respecto T. Graham, numerosos aspectos de conformidad Unicode que XML no respetaba. Para una mayor información sobre estas diferencias así como sobre la evolución sufrida por XML para adaptarse al estándar Unicode, véase Tony Graham. *Changes in Unicode that led to changes in XML 1.0 Second Edition* [documento HTML]. IBM, January 2001. Disponible en <http://www-106.ibm.com/developerworks/library/u-xml.html?dwzone=xml> (consultado el 31 de enero de 2001).

- Toda etiqueta inicial debe tener su correspondiente etiqueta final, excepto en el caso de que se trate de un elemento vacío; por ejemplo:

```
<documento> . . . . </documento>
```

- En el caso de tratarse de un elemento vacío es necesario indicar este hecho mediante la inclusión de la barra oblicua antes del ángulo de cierre de la etiqueta inicial aunque también es correcto sintácticamente utilizar el par de etiquetas habituales sin encerrar, en este caso, contenido alguno entre ellas (menos recomendable); por ejemplo:

```
<br/> o también <br></br>
```

- El nombre que se emplee para los nombres de los elementos (identificadores genéricos) ha de ser idéntico en la etiqueta inicial y final siendo, además, estos nombres son sensibles al tipo de letra (mayúsculas, minúsculas, acentos, etc.); por ejemplo:

```
<capitulo> . . . </Capitulo>  
sería incorrecto pues debería ser  
<capitulo> . . . . </capitulo> o <Capitulo>...  
</Capitulo>
```

- Todos los elementos deben estar anidados de forma adecuada; por ejemplo:

```
<capitulo><apartado>....</apartado></capitulo>
```

- Un atributo no puede aparecer más de una vez dentro del mismo elemento y, además, todos los valores de los atributos deberán ir siempre entre comillas. Por otro lado, el valor de un atributo nunca podrá ser una referencia a una entidad externa. Un ejemplo de atributo-valor en XML sería, por ejemplo:

```
<fecha tipo="num">..... </fecha>
```

- Los caracteres de texto &, <, >, ‘, “ deben ser siempre representados como referencias a entidades de carácter dado que son consideradas como entidades predefinidas y, por tanto, interpretadas con un significado especial por cualquier procesador XML; por ejemplo:

```
<empresa>Cosméticos Asociados &amp; Cia.</empresa>
```

Por el contrario, un **documento válido** es aquel que siendo correcto sintácticamente se declara conforme a las leyes establecidas por una DTD referenciada a través de la correspondiente declaración de tipo de documento<sup>114</sup>. El modelo de documento XML válido (el único posible para el metalenguaje SGML) es utilizado para aquellas aplicaciones en las cuales resulta necesario verificar que la estructura lógica específica de un documento se ajusta a la estructura lógica genérica del tipo documental al cual se adscribe. O dicho de otro modo, un documento XML válido conlleva que la instancia del documento XML se ajusta a los modelos de contenidos de los elementos y a los tipos de valores de éstos y de sus atributos definido todo ello en cada una de las declaraciones existentes en la DTD, como ya se comentó ampliamente en el capítulo anterior en el apartado correspondiente al uso de la DTD en documentos SGML.

El hecho de la existencia de estos dos tipos de documentos dentro del metalenguaje XML, es decir, documentos XML con y sin definición formal de tipo documental, ha sido desde su origen un tema ciertamente polémico que ha ocasionado ríos de tinta entre los investigadores de este campo. En la raíz de este problema subyace el propio origen del mismo (desde el mundo de la informática) y las aplicaciones principales hacia las que se orientó desde un principio el empleo de este metalenguaje (tratamiento de datos en el entorno Web, principalmente). Aunque este tema será tratado con mayor detalle en próximos apartados de este capítulo es necesario hacer aquí una primera aproximación al tema, señalando brevemente las ventajas y desventajas de cada uno de estos modelos. Para ello nos remitimos a lo expuesto por M. Maziarka (de la compañía norteamericana Xyvision) en la conferencia europea sobre SGML y XML celebrada en París en 1998<sup>115</sup>:

- Las ventajas que se derivan del empleo de documentos XML válidos están en el hecho de que al definir previamente las estructuras de los documentos los datos a incluir en

---

<sup>114</sup> Natanya Pitts-Moultis, Cheryl Kirk. *XML Black Book*. Scottsdale, Arizona: Coriolis Group, 1999, p. 26.

<sup>115</sup> Michael Maziarka. "XML is not just another name for SGML. XML is the vehicle to deploy structured data systems throughout an organization" [documento HTML]. En: *SGML/XML Europe 98 Conference* (1998. Paris). GCA, 1998. Disponible en <http://www.infoloom.com/gcaconfs/WEB/paris98/maziarka.HTM> (consultado el 5 de octubre de 2000).



los nuevos documentos seguirán unas pautas y reglas ya definidas para todos ellos por lo que se producen menos errores en dicha generación de documentos y se garantiza una normalización o uniformidad de dichos datos. Asimismo, las aplicaciones informáticas que han de procesar estos datos pueden ser programadas con anterioridad basándose en la clasificación, tipología y orden de aparición de los datos en los documentos XML válidos facilitando, por tanto, el procesamiento de dichos datos. Por el contrario, las desventajas que se pueden encontrar en este modelo apuntan a la complejidad del proceso de definición de estructuras documentales, siendo éste un trabajo arduo que implica unas inversiones iniciales elevadas en tiempo de dedicación y en profesionales cualificados en este campo. Asimismo, para los creadores de los datos y para los editores supone tener que trabajar con unas restricciones de uso establecidas previamente.

- En el caso del modelo constituido por los documentos bien formados sus beneficios se desprenden de la reducción o eliminación de los costes asociados al desarrollo de una o varias DTDs así como de la libertad que se les proporciona a los autores para crear estructuras de datos no normalizadas pero adecuadas para representar la información que se desea procesar. Los inconvenientes de construir estructuras *ad hoc* para documentos XML están en el hecho de que el procesamiento posterior de los datos contenidos en dichos documentos puede resultar más difícil y complejo y, además, estos datos no podrán ser reutilizados fuera de un contexto documental predefinido.

Por último, para tener un conocimiento exacto del alcance y prestaciones del metalenguaje XML resulta necesario detallar brevemente las principales diferencias existentes con respecto al mecanismo existente en el metalenguaje SGML para elaborar definiciones de tipos documentales (DTDs) pues es aquí donde se encuentran el grueso de la reducción de esta norma internacional. De forma general, la mayor parte de estas restricciones y supresiones están relacionadas con los mecanismos disponibles para definir tipos documentales o DTDs (lenguajes específicos de marcado) que, como sabemos, en el estándar internacional SGML ocupaba un papel principal. Por contra, XML elimina más de 30 construcciones utilizadas en SGML para definir sus DTDs que, bien, se utilizaban poco

o, bien, introducían ambigüedades en su interpretación haciendo difícil y complejo su tratamiento por parte de los programas informáticos<sup>116</sup>. Expondremos brevemente algunas de esas diferencias<sup>117</sup>:

- En SGML, aquellos elementos que incluyen datos de carácter que no han de ser analizados por la aplicación (CDATA) establecían esta circunstancia en su correspondiente declaración dentro de la DTD. XML, por el contrario, requiere que el contenido CDATA sea representado de forma especial en la instancia del documento, no en la DTD, del modo siguiente:

```
<etiqueta>
<![CDATA[
...datos de carácter no analizados... ]]>
</etiqueta>
```

- Si se hace uso de los comentarios dentro de la DTD, éstos no podrán ir dentro de las declaraciones por lo que tendrán que producirse de forma aislada, como se muestra en el ejemplo:

```
<!ELEMENT fecha EMPTY>
<!-- se definen a continuación los atributos para este
elemento -->
<!ATTLIST fecha .....
```

- No está permitido realizar declaraciones conjuntas de grupos de elementos o de atributos por lo que cada elemento o atributo debe ser definido de forma independiente a través de su correspondiente declaración; por ejemplo:

No es válido

```
<!ELEMENT (nombre | apellido1 | apellido2) (#PCDATA)>
Y sí lo es
<!ELEMENT nombre (#PCDATA)>
```

---

<sup>116</sup> T. A. Powell. *Op. cit.*, p. 614.

<sup>117</sup> Además del anteriormente citado artículo de J. Clark, publicado como “Nota” dentro del W3C, resulta de gran interés el artículo elaborado por la compañía alemana STEP (*Stürt Electronic Publishing*) dada su claridad expositiva, por lo que recomendamos su consulta para una mayor profundidad en este tema. Véase en *Introduction to XML* [documento HTML]. STEP, [1999]. Disponible en [http://www.oasis-open.org/html/step\\_intro\\_to\\_xml.html](http://www.oasis-open.org/html/step_intro_to_xml.html) (consultado el 9 de diciembre de 1999).

```
<!ELEMENT apellido1 (#PCDATA)>  
<!ELEMENT apellido2 (#PCDATA)>
```

- La opcionalidad en la reducción o minimización de las etiquetas existente en el modelo de contenido de elementos de la DTD de SGML queda suprimida, por lo que todo elemento cuyo modelo de contenido en XML no esté definido deberá llevar sus correspondientes etiquetas inicial y final en la instancia del documento, como ya se había comentado con anterioridad. Este hecho conlleva, por tanto, que en la definición de los elementos en la DTD de XML no aparecerán los signos “-“ o “0”, indicativos de la posibilidad de aparición o no de dichas etiquetas; por ejemplo:

No es válido

```
<!ELEMENT nombre - 0 (#PCDATA)>
```

Y sí lo es

```
<!ELEMENT nombre (#PCDATA)>
```

- XML elimina el uso de las excepciones al modelo de contenido de los elementos por lo que no son admitidas las exclusiones ni las inclusiones de otros elementos no englobados entre los paréntesis del modelo de contenido.

No es válido

```
<!ELEMENT seccion - - (titulo, parrafo+) +(nota)>
```

Y sí lo es

```
<!ELEMENT seccion (titulo | parrafo | nota)*>
```

- No se da soporte para el operador AND (definido con el *ampersand*) dentro del modelo de contenido existiendo, por tanto, únicamente la secuencia de elementos (definido con la coma) o la opcionalidad en la aparición de uno de ellos (definido con la barra vertical).
- Si un elemento viene definido con un modelo de contenido mixto, el valor de #PCDATA deberá figurar en primer lugar y el resto de elementos contenidos deberán ser separados con la barra vertical de opcionalidad. De igual forma, el signo del asterisco tendrá que ser empleado como único indicador de frecuencia posible; por ejemplo:

<!ELEMENT parrafo (#PCDATA | referencia)\*>

- El nombre reservado de RCDATA en las declaraciones de elementos para señalar el procesamiento de un tipo de datos especiales no existe.
- Dentro de la declaración de atributos de elementos XML se suprimen los valores de atributos NAME(S), NUMBER(S) y NUTOKEN(S). De igual forma, se suprimen los valores CURRENT y CONREF para indicar el comportamiento del valor por defecto definido para un atributo.
- Las entidades SDATA para la definición de caracteres especiales no existen.
- Las entidades externas tienen que ser declaradas mediante un identificador de sistema (SYSTEM) y, como opción adicional, a través de un identificador público (PUBLIC). El identificador de sistema deberá ser un URI, que en la mayor parte de los casos será un URL o nombre de ruta hacia el fichero en cuestión.

Anotar, finalmente, que esta importante reducción del SGML en XML hace que se pierda una parte importante de la capacidad y potencialidad existente con SGML para el tratamiento de documentos electrónicos. De igual modo, se pierde la rigurosidad restrictiva de las DTD SGML aumentando la ambigüedad en algunos modelos de contenidos de elementos en las DTDs XML. Al contrario de lo señalado por Bartlett, no creemos posible que algunas de estas capacidades omitidas sean repuestas en futuras revisiones dentro del propio metalenguaje<sup>118</sup> dado que el aumento en la potencialidad de XML está viniendo de la mano de los lenguajes asociados al mismo, los cuales complementan y potencian aquellos aspectos mejorables de este modelo, como es el caso de los esquemas XML, la definición de enlaces hipertextuales o la asignación de estilos de presentación de los documentos, como se expondrá en apartados posteriores de este capítulo.

---

<sup>118</sup> P.G. Bartlett. *Op. cit.*, <http://www.infoloom.com/gcaconfs/WEB/paris98/bartlett.HTM>

### III.2.2.3. BENEFICIOS E INCONVENIENTES

Dado que XML deriva directamente del metalenguaje SGML proporcionará, por tanto, beneficios similares a los descritos en el anterior capítulo para dicho metalenguaje. Pero dada la doble orientación con la que fue creado XML, esto es, a documentos y a datos en un entorno de trabajo Web, como se analizará en apartados posteriores de este capítulo, este nuevo metalenguaje obtendrá otras ventajas añadidas que no existían en el estándar internacional SGML.

Entre los numerosos beneficios que se desprenden del empleo del metalenguaje XML por parte de cualquier organización para la generación y tratamiento de documentos electrónicos, podemos destacar los siguientes<sup>119</sup>:

- **XML es un estándar abierto:** Como ya se ha expuesto, el *World Wide Web Consortium* (W3C) vela por los intereses en el desarrollo y mantenimiento de este estándar para la industria. Si bien es cierto, como ya se comentó con anterioridad, que el W3C no posee oficialmente la categoría de institución supranacional de normalización como sería la ISO, el apoyo recibido por parte de múltiples instituciones, organizaciones y grandes compañías informáticas así como de los usuarios de la Web convierten a sus “recomendaciones” en estándares internacionales *de facto*. Así, tratando de conjugar los intereses de todas las partes el W3C consolida estándares en un entorno de trabajo

---

<sup>119</sup> El siguiente listado de beneficios derivados del metalenguaje XML, así como los principales obstáculos con los que se encuentra esta nueva tecnología, han sido extraídos de lo expuesto en diversos artículos y manuales de interés ya referenciados (o por referenciar), entre los que podemos destacar los siguientes:

- J. Sturm. *Op. cit.*, p. 21.
- Nazmul Idris. *Benefits of using XML* [documento HTML]. Developerfile.com, June 13, 1999. Disponible en <http://www.developerfile.com/xmlbenefits/default.htm> (consultado el 23 de junio de 2000).
- Adrian Rivers. “XML (Extensible Markup Language): If we ignore it, will it go away?”. *Information Management & Technology*, v. 33, n° 1, 2000, p. 39.

abierto a cualquier institución, empresa o particular que desee participar en sus desarrollos normativos.

- **XML es independiente de la plataforma y del sistema operativo:** Al igual que su predecesor, el SGML, XML está diseñado de tal modo que los documentos producidos son almacenados en texto plano. Este hecho proporciona que dichos documentos puedan ser procesados en cualquier plataforma informática y por cualquier sistema operativo existente hoy en día en el mercado. Asimismo, esta independencia física y lógica hace que los documentos XML sean más robustos y estables en el tiempo.
- **XML es un metalenguaje internacional:** A diferencia de SGML y HTML que estaban basados en la codificación de caracteres ASCII, XML está perfectamente preparado para trabajar con caracteres Unicode. Este sistema de codificación permite incluir caracteres de cualquiera de los alfabetos humanos (y no humanos) empleados en todo el mundo.
- **XML es completamente extensible:** al no ser XML un lenguaje de marcado predefinido y sí un metalenguaje, proporciona a los usuarios la capacidad para que éstos definan sus propias etiquetas de marcado de texto según sean sus necesidades específicas. Este hecho permite bien, definir *ex novo* un vocabulario de marcado para nuestros documentos o, bien, trasladar estructuras y nombres derivados de bases de datos ya existentes en la organización.
- **XML define estructuras lógicas y físicas:** Cuando marcamos el texto de un documento mediante XML estamos reflejando implícitamente las relaciones estructurales que se dan entre esos bloques o piezas marcadas, estableciendo un orden de jerarquía y subordinación entre las mismas. Además, a través de la DTD se

---

- Deborah S. Ray, Eric J. Ray. "Bridging the Gap Between SGML and HTML: The Potential of XML for Technical Communicators". *Technical Communication*, v. 45, n° 3, 1998, p. 431.

proporciona un mecanismo para la definición formal de estructuras documentales similares entre grupos de documentos, haciendo posible la creación de documentos XML normalizados y válidos. De igual forma, si éstos están compuestos físicamente por entidades, es posible definir fácilmente su estructura física.

- **XML se orienta tanto a documentos como a datos:** Como se expondrá con mayor profundidad en el apartado dedicado a las orientaciones en la aplicación de XML, este metalenguaje está especialmente diseñado para el procesamiento de documentos electrónicos (capacidad heredada de SGML) así como para el procesamiento de datos (capacidad heredada de su proximidad conceptual con las bases de datos). Esta última propiedad permite a este estándar ser considerado como el mejor modelo actual para el intercambio de datos a través de redes informáticas, dada la gran capacidad de los documentos XML para contener fragmentos de datos.
  
- **XML no sólo se orienta a la WWW:** Aunque en un principio XML fue diseñado para su implementación en la *World Wide Web*, esta tecnología está preparada para ser utilizada en cualquier tipo de red informática y de telecomunicaciones, tanto de área extensa como de área local, que utilice la tecnología, protocolos y servicios de la red Internet. De hecho, gran parte del potencial de la tecnología XML se adapta mejor al trabajo en redes corporativas o intranets que al macroespacio de la Web dado que aquí se puede hacer un mejor control y gestión de los documentos y datos procesados mediante la tecnología XML.
  
- **XML facilita la búsqueda y recuperación de información:** Se trata éste de uno de los aspectos más interesantes de este metalenguaje dado que ahora, y a diferencia del lenguaje HTML, se pueden describir con exactitud la ubicación de las piezas informativas de interés dentro de los documentos electrónicos en un contexto estructural y semántico de producción. Este hecho es posible gracias a la identificación *per se* de dichas piezas informativas marcadas con las correspondientes etiquetas

identificativas o, bien, al ser posible la asociación de mecanismos para la definición de metainformación (metadatos) que caracterizan al documento.

- **XML es una tecnología sencilla de usar y económica de implementar:** En principio, y para un uso básico, aprender a confeccionar documentos XML no resulta especialmente difícil de aprender dado que las reglas sintácticas tanto para la construcción de documentos como para la definición de tipos documentales son más reducidas y sencillas que las existentes para SGML. La complejidad de XML vendrá principalmente por la aplicación del resto de estándares asociados al mismo, como se explicará en el siguiente apartado. De igual modo, las herramientas informáticas capaces de procesar y tratar documentos XML son más económicas que las existentes para SGML (de hecho ésta fue una de las premisas básicas de partida en la creación de XML). Así, gran parte del software existente para SGML se ha adaptado para el trabajo con este tipo de documentos (por ejemplo, los editores gráficos o los gestores de bases de datos) junto con otros muchos que están apareciendo creados específicamente para trabajar en este entorno que, o bien son gratuitos o bien tienen unos precios bastante asequibles para cualquier institución.
  
- **XML facilita el establecimiento de estándares para todo tipo de datos:** A través del metalenguaje XML es posible tratar cualquier tipo de documento o de datos lo que posibilita el establecimiento de diversos estándares por parte de las organizaciones y empresas que los manejan. Así, toda aquella información que sea susceptible de ser estructurada es materia del metalenguaje XML, por lo que es posible encontrar vocabularios normalizados para el tratamiento de gráficos vectoriales, fórmulas matemáticas, notaciones musicales, o cualquiera de los tipos documentales clásicos que se manejan en el quehacer cotidiano de las organizaciones, como se explicará en el capítulo III.3 de esta tesis.

Si bien es verdad que con esta nueva tecnología son innumerables los beneficios que se pueden obtener en el tratamiento y gestión de la información electrónica dentro de

cualquier institución, no es menos cierto que existen algunas **limitaciones** o desventajas que hacen que su aplicación sea hoy en día tomada con cierto recelo. Entre los principales obstáculos con los que la tecnología derivada de XML se encuentra destacaríamos los siguientes:

- **XML es todavía una tecnología inmadura**<sup>120</sup>: XML es, a pesar de los años transcurridos, una tecnología aún novedosa donde una parte de los estándares que lo complementan y potencian no están aún perfectamente asentados (algunos de ellos se encuentran aún en fase de elaboración o han sido recientemente aprobados como recomendación del W3C). Este hecho conlleva que las herramientas informáticas que han de procesar los documentos XML se encuentren igualmente en fase de desarrollo y muchas de ellas sean incapaces de tratar de forma adecuada y completa la tecnología asociada a XML (por ejemplo, muchos de los navegadores web son incapaces hoy en día de trabajar con los enlaces hipertextuales extendidos para XML)<sup>121</sup>.
- **XML conlleva nuevos riesgos de inversión económica y de formación profesional**: Muchas instituciones recelan de los beneficios publicitarios que invariablemente suelen venir asociados con la aparición de cada nueva tecnología informática, más aún cuando ello conlleva realizar nuevas inversiones económicas tanto

---

<sup>120</sup> Existen, además de lo aquí expuesto, otros aspectos en esta inmadurez tecnológica por parte de XML que no van a ser tratados en esta tesis doctoral debido a que hacen mención a las cuestiones puramente técnicas del funcionamiento de la red Internet (protocolos, identificadores de contenido MIME, URIs, infraestructuras y seguridad, etc.) y que, por tanto, salen fuera de la intencionalidad de esta tesis. En cualquier caso, sí resulta recomendable la consulta del artículo de Simon St. Laurent. *XML: A Disruptive Technology* [documento HTML]. XML.com, June 21, 2000. Disponible en <http://www.xml.com/print/2000/06/21/disruption/index.html> (consultado el 5 de julio de 2000) para profundizar en estos otros aspectos de desarrollo y evolución tecnológica que trae asociados esta nueva tecnología en la esfera de la gran red de comunicación que es la Internet.

<sup>121</sup> A pesar de que siempre es aventurado realizar “estados de la cuestión” en el mundo de los programas informáticos debido a la continua evolución y constante renovación en este campo, resulta recomendable la consulta del documento de S. St. Laurent (y documentos asociados al mismo) para conocer con exactitud las posibilidades reales de trabajo con documentos XML de los tres principales navegadores Web existentes hoy en día: Mozilla/Netscape, Opera e Internet Explorer. Véase Simon St. Laurent. *Browser XML Display Support Chart* [documento HTML]. XML.com, May 2, 2000. Disponible en <http://www.xml.com/print/2000/05/03/browserchart/index.html> (consultado el 15 de junio de 2000).

en infraestructuras tecnológicas como en formación del personal. De igual modo, en muchas de estas instituciones, tras un proceso, a veces largo y costoso, de asentamiento y consolidación de las tecnologías informáticas para el tratamiento de la información electrónica generada, los profesionales encargados de dichas tareas documentales sienten cierta inseguridad en experimentar con otras tecnologías de las cuales pueden tener escasas o nulas referencias. Los procesos formativos para adaptarse al trabajo con los documentos XML pueden ser a veces largos y complejos pues ello trae consigo una nueva mentalidad en la filosofía de trabajo dentro de las organizaciones, no siempre fácil de asumir por parte de todos los empleados<sup>122</sup>.

- **XML se basa en el marcado de los documentos y la definición de tipos documentales:** A diferencia de otros sistemas de edición electrónica, XML exige el marcado de las piezas estructurales y semánticas de los documentos, labor ésta que, aunque de vital importancia, resulta a menudo ciertamente compleja, lenta y tediosa. A pesar de que día a día están apareciendo en el mercado nuevas herramientas informáticas que simplifican estas tareas no es menos cierto que el marcado del texto siempre consumirá más tiempo y concentración que la simple utilización de los habituales y rápidos procesadores de texto WYSIWYG (MSWord, WordPerfect, etc.) que se manejan en la mayoría de las instituciones. De igual forma, el empleo de la tecnología XML conllevará en la mayoría de los casos la definición formal *ex novo* de tipos documentales existentes en la organización a través del modelo de DTD o de esquema XML (o la adaptación de algunas ya existentes). Esta labor requiere profesionales expertos en la materia, siendo, igualmente, una labor lenta y compleja de realizar.
- **XML fomenta las guerras por el establecimiento de estándares:** XML, como metalenguaje que es, permite crear lenguajes específicos para la transmisión e

---

<sup>122</sup> Esta importante limitación u obstáculo es señalado perfectamente, aunque rebatido por los beneficios obtenidos como contrapartida, por E. Willner en la primera parte del artículo dedicado a explicar los usos y beneficios obtenidos en la utilización de los metalenguajes SGML y, especialmente, XML. Véase Eli Willner. "Preparing Data for the Web with SGML/XML". *Information Today*, v. 15, n° 5, May 1998, p. 54.

intercambio de datos y documentos electrónicos entre ordenadores. El problema radica en quién ha de establecer estos lenguajes específicos que se han de convertir en estándares para los distintos sectores de producción documental. Este hecho fomenta que existan muchas compañías, empresas y organizaciones trabajando, en algunos casos, de forma aislada y compitiendo entre sí para imponer su modelo de lenguaje de marcado de documentos. El caso más notorio se da, lógicamente, en aquellos sectores en los que entran en juego grandes intereses económicos, y de forma muy especial en todo lo relacionado con el comercio electrónico en Internet. En este contexto compiten, como es señalado por S. M. Dugan, múltiples estándares XML, entre los que destacan el desarrollo BizTalk de las compañías Microsoft, IBM y otras tantas asociadas y el estándar ebXML desarrollado por la compañía Sun Microsystems<sup>123</sup>. Esta misma idea se puede establecer a otros muchos sectores de producción y tratamiento documental donde no siempre es fácil poner de acuerdo a todos los interlocutores, cada uno con unos determinados intereses particulares.

---

<sup>123</sup> Para una mayor información sobre este tema, véase Sean M. Dugan. *The Web's search for meaning* [documento HTML]. ComputerUser.com, March 1, 2001. Disponible en <http://www.computeruser.com/articles/2003,1,3,1,0301,01.html> (consultado el 14 de junio de 2001).

### III.2.3. ESTÁNDARES XML ACOMPAÑANTES<sup>124</sup>

XML, como ya se ha comentado, fue concebido en su origen como un núcleo inicial pequeño pero estable al cual se le irían asociando con el tiempo otros estándares que lo complementarían y potenciarían con diferentes funcionalidades. Por esta razón, XML es algo más que un simple lenguaje (o metalenguaje) de marcas para la Web: se trata de un amplio abanico de tecnologías asociadas que, junto a la especificación primitiva, conforman una gran familia de herramientas de trabajo interrelacionadas. Esta concepción permite a las aplicaciones desarrolladas al amparo de XML obtener una potencia y versatilidad elevadas en el tratamiento de información electrónica de cualquier tipo<sup>125</sup>.

El W3C ha venido desarrollando un gran número de estándares asociados al metalenguaje XML, a los cuales se les ha venido denominado “estándares XML acompañantes” y que son, en realidad, los que hacen que el estudio e implementación de soluciones XML potentes sea una labor altamente especializada y compleja. Algunos de estos estándares no lo son tal pues aún están en fase de desarrollo y normalización y, en otros casos, se trata únicamente de la aplicación de tecnologías desarrolladas en un principio para el metalenguaje SGML o el lenguaje HTML y que se han ido adaptando a la sintaxis concreta del metalenguaje XML.

Resulta, pues, necesario exponer aquí, aunque de forma breve dado que esta tesis no pretende ser un “manual de uso”, el funcionamiento de aquellos mecanismos asociados a XML en los que se apoyará principalmente el modelo aquí propuesto para el tratamiento de la documentación electrónica jurisprudencial emanada del Tribunal Constitucional español

---

<sup>124</sup> Debido a los rápidos cambios que se producen en el seno del W3C con relación a la situación en la fase normativa en la que se encuentran los documentos elaborados por los distintos grupos de trabajo, recomendamos la consulta permanente de la página web de Berners-Lee correspondiente a dicha materia, y citada con anterioridad, en la dirección <http://www.w3.org/TR/> De igual modo, y para obtener una visión básica y general de los estándares acompañantes al metalenguaje XML recomendamos la lectura del documento de Lisa W. Bos. “Overview of XML Family of Standards” [documento HTML]. En: *XML 99 Conference* (Philadelphia, December 1999). GCA, 1999. Disponible en <http://www.infoloom.com/gcaconfs/WEB/philadelphia99/lisabos.HTM> (consultado el 22 de enero de 2001).

<sup>125</sup> B. Marchal. *Op. cit.*, p. 32.

y presentado en el siguiente capítulo. Así, entre las piezas que componen el complejo puzzle de la tecnología XML se pueden destacar las siguientes: la definición de estructuras lógicas de los tipos documentales manejados y los modelos existentes en la actualidad para dicha definición (DTD y XML Schema, los cuales tendrán un amplio tratamiento en apartados posteriores); los mecanismos existentes para el establecimiento de potentes vínculos hipertextuales entre los documentos relacionados explícitamente (XLink y XPointer); los modelos asociados a los documentos XML para asignar diversos formatos de presentación según sea el medio de salida seleccionado (CSS y XSL) así como los mecanismos para la transformación de los documentos XML a otros formatos electrónicos de publicación ampliamente difundidos (XSLT); los mecanismos existentes para la correcta descripción de la metainformación contenida en el documento XML (RDF, el cual tendrá un amplio tratamiento en apartados posteriores de este capítulo); las diversas opciones para el almacenamiento, localización y recuperación de información existente en la base de datos XML construida a tal fin (XML Query), o, por último, el nuevo modelo planteado para el desarrollo de formularios Web (XForms).

Algunas de estas tecnologías se encuentran aún, como decíamos con anterioridad, en fase de desarrollo por lo que su implementación práctica por parte del software existente para la edición, almacenamiento, procesamiento, localización y presentación de documentos XML resulta, hoy por hoy, ciertamente limitada, como se irá exponiendo en apartados posteriores de esta tesis cuando se tenga que hacer uso de ellas para el desarrollo práctico del modelo realizado.

En cualquier caso, a continuación se describen brevemente estos estándares acompañantes, destacando por su importancia y desarrollo a los siguientes:

- **XLink (XML Linking Language)**<sup>126</sup>: Fuertemente influenciado en su origen por el estándar internacional HyTime de la ISO, el lenguaje de enlace para XML vino en un principio descrito, como ya se comentó con anterioridad, como parte integrante de los primeros borradores que se confeccionaron sobre el metalenguaje XML, del cual se segregaría con posterioridad para tener su propio desarrollo normativo dentro del W3C. A pesar de este temprano desarrollo XLink ha sido aprobado como recomendación del W3C en fechas recientes (recomendación de 27 de junio de 2001<sup>127</sup>). XLink define diversos mecanismos para establecer las vinculaciones o relaciones hipertextuales entre documentos XML, ampliando espectacularmente las capacidades de los sencillos enlaces hipertextuales unidireccionales existentes para el lenguaje HTML. Al igual que en HTML se definen enlaces simples entre dos documentos pero XLink añade la capacidad de crear vínculos extendidos entre documentos, siendo posible especificar las direcciones de más de un enlace de destino. De igual forma, este lenguaje posibilita componer documentos XML mediante la vinculación de los elementos que se van a insertar dentro del mismo. Asociado a XLink existen otros lenguajes que lo complementan en sus capacidades, como son el caso, entre otros, de XPointer, XPath y XML Base. **XPointer (XML Pointer Language)**<sup>128</sup>, aún en fase de borrador de trabajo, es el lenguaje que permite aumentar la precisión en las descripciones de recursos enlazados mediante XLink facilitando la localización exacta de un determinado recurso dentro de un documento XML. Para llevar a cabo dicha labor, este lenguaje utiliza punteros XML que se incluyen dentro del URL al cual señala el enlace hipertextual definido en el documento XML. Para la ayuda en la localización exacta del recurso dentro de la

---

<sup>126</sup> Toda la información oficial relativa al lenguaje de enlaces hipertextuales XML y demás lenguajes asociados a éste se encuentra disponible en la dirección web del W3C en <http://www.w3.org/XML/Linking/>, documento creado en origen por Dan Connolly y mantenido en la actualidad por Daniel Veillard.

<sup>127</sup> World Wide Web Consortium. *XML Linking Language (XLink) Version 1.0. W3C Recommendation* [documento HTML]. Steve DeRose, Eve Maler, David Orchard (eds.). W3C, June 27, 2001. Disponible en <http://www.w3.org/TR/xlink/> (consultado el 27 de junio de 2001).

<sup>128</sup> World Wide Web Consortium. *XML Pointer Language (XPointer) Version 1.0. W3C Last Call Working Draft* [documento HTML]. Steve DeRose, Eve Maler, Ron Daniel (eds.). W3C, January 8, 2001. Disponible en <http://www.w3.org/TR/xptr/> (consultado el 16 de febrero de 2001).

estructura de elementos y atributos existentes en un documento XML la especificación XPointer se ha venido basando en la sintaxis definida para tal fin en otro desarrollo anterior del consorcio, conocido como **XPath** (*XML Path Language*)<sup>129</sup>, recomendación del W3C desde noviembre de 1999. Por último, otro lenguaje íntimamente asociado a XLink es el lenguaje denominado **XML Base**<sup>130</sup>, recomendación del W3C desde junio de 2001, el cual proporciona una sintaxis propia para realizar la misma funcionalidad que tiene el elemento BASE dentro de los documentos HTML; esto es, indicar explícitamente la base del URL de todos los enlaces hipertextuales que se vayan a incluir en el documento.

- **XML Namespaces**<sup>131</sup>: Los “nombres locales XML”<sup>132</sup>, recomendación del W3C desde enero de 1999, fueron creados para implementar la función CONCUR de SGML que no había sido prevista (y, por tanto, no incluida) en la especificación de XML. Esta importante función permitía que un documento SGML pudiera contener de forma

---

<sup>129</sup> World Wide Web Consortium. *XML Path Language (XPath) Version 1.0. W3C Recommendation* [documento HTML]. James Clark, Steve DeRose (eds.). W3C, November 16, 1999. Disponible en <http://www.w3.org/TR/xpath/> (consultado el 16 de febrero de 2001).

<sup>130</sup> World Wide Web Consortium. *XML Base. W3C Recommendation* [documento HTML]. Jonathan Marsh (ed.). W3C, June 27, 2001. Disponible en <http://www.w3.org/TR/xmlbase/> (consultado el 27 de junio de 2001).

<sup>131</sup> World Wide Web Consortium. *Namespaces in XML. W3C Recommendation* [documento HTML]. Tim Bray, Dave Hollander, Andrew Layman (eds.). W3C, January 14, 1999. Disponible en <http://www.w3.org/TR/REC-xml-names/> (consultado el 9 de diciembre de 2000). Asimismo, y dado que el tema de los *XML Namespaces* no se volverá a tocar más en esta tesis doctoral, recomendamos para una más amplia información sobre esta materia el excelente documento de R. Bourret, disponible en Ronald Bourret. *XML Namespaces FAQ* [documento HTML]. RPBOURRET.com, rev. November 2000. Disponible en <http://www.rpbouret.com/xml/NamespacesFAQ.html> (consultado el 31 de enero de 2001).

<sup>132</sup> Resulta frecuente encontrar en diversos manuales y documentos sobre XML traducidos o elaborados en nuestro idioma el empleo de “espacios de nombre XML” como traducción literal del término anglosajón. Consideramos más precisa y ajustada la traducción que realiza J. A. Sordo Díaz al denominar a esta recomendación bajo los términos de “nombres locales XML” pues, como se verá a continuación, se trata de un mecanismo normalizado para identificar y diferenciar dentro de un documento XML los nombres de elementos y atributos que se han establecido de forma “local” por alguna institución frente a otros con nombres idénticos. Para una mayor información, véase José Alberto Sordo Díaz. *XML Namespaces* [documento MSWord]. Getafe: Universidad Carlos III de Madrid, 3 de julio de 2000. Disponible en <http://xml.it.uc3m.es/documentos/traduccion/namespaces.doc> (consultado el 21 de enero de 2001).

simultánea etiquetas provenientes de distintas DTDs. Así, mediante el mecanismo de los *namespaces* se posibilita que una instancia de documento XML pueda contener elementos procedentes de distintas DTDs, identificando y distinguiendo los nombres de elementos y atributos de cada uno de los lenguajes específicos empleados dentro de un mismo documento XML. De forma similar, este mecanismo posibilita una correcta distinción de etiquetas cuando dentro de un documento XML se están empleando nombres de elementos y/o atributos coincidentes procedentes de diversas DTDs. De este modo se consigue que cada nombre de elementos y/o atributo empleado dentro de un documento XML sea único y universal. Para llevar a cabo esta función la especificación de los nombres locales XML proporciona un mecanismo sencillo para cualificar cada nombre de elemento y/o atributo asignando para ello un sistema de nombres dividido en dos partes para cada elemento y atributo que lo requiera: la primera parte del nombre viene constituida por el URI empleado para identificar ese nombre local (*the namespace name*), y la segunda parte viene constituida por el nombre del elemento o atributo en sí mismo (*the local part*). Para declarar el uso de nombres locales dentro de documentos XML se ha de hacer uso del atributo “*xmlns:prefijo*” y con valor igual al URI en el que se establece el lenguaje del cual se están utilizando sus elementos y/o atributos, situando normalmente dicho atributo y su valor dentro del elemento raíz. Los nombres locales XML se han venido empleando con asiduidad para, por ejemplo, hacer uso dentro de los documentos XML de aquellos elementos y atributos procedentes del lenguaje HTML que permitían de forma rápida y directa obtener ciertas respuestas de visualización y comportamiento por parte de los navegadores web, como es el caso de la inserción de imágenes, el establecimiento de anclas para las relaciones hipertextuales entre documentos, la incorporación de formularios, etc. El siguiente ejemplo ilustra perfectamente el modo de introducir una imagen dentro de un documento XML haciendo uso para ello de la referencia al nombre local del elemento correspondiente en el lenguaje HTML:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<agenda xmlns:html="http://www.w3.org/TR/REC-html40">
<html:img src="boni.gif"/>
<nombre>Bonifacio</nombre>
<apellido_1>Martín</apellido_1>
```

```
<apellido_2>Galán</apellido_2>
<html:a
href="mailto:bmartin@bib.uc3m.es">bmartin@bib.uc3m.es</html:
a>
<html:a
href="http://www.bib.uc3m.es/~bmartin/">http://www.bib.uc3m.
es/~bmartin/</html:a>
</agenda>
```

- **XSL (XML Stylesheet Language):** La asignación de estilo de presentación en el entorno de trabajo viene asociado al lenguaje extensible de estilo XML aunque, igualmente, los documentos XML permiten la inclusión de hojas de estilo CSS, lenguaje de estilo asociado en origen al HTML y ya comentado en apartados anteriores, pero con serias limitaciones para un uso efectivo en este nuevo contexto tecnológico<sup>133</sup>. Aunque XSL fue planteado en un principio como un único lenguaje para el formateado de documentos XML, fuertemente influenciado por el lenguaje DSSSL (*Document Style Semantics and Specification Language*) para documentos SGML, en la actualidad este lenguaje hace referencia a un conjunto normativo organizado, en realidad, en dos partes o dos lenguajes específicos con desarrollos paralelos y con independencia funcional: uno orientado a establecer mecanismos para la transformación de documentos XML (a otros formatos o en otro documento XML distinto) y otro orientado a la asignación de un formato de presentación o de estilo, propiamente dicho, de documentos XML.

---

<sup>133</sup> Sin entrar en un profundo detalle de las limitaciones del lenguaje CSS frente al nuevo lenguaje XSL para la asignación de uno varios estilos de presentación a los documentos XML, sí resulta conveniente apuntar mínimamente dichas limitaciones. Además de ser un lenguaje no escrito con la sintaxis propia de XML, las principales limitaciones se encuentran, como así es señalado por P. Kamthan, en las siguientes: CSS no puede tomar un ítem desde un lugar y utilizarlo con posterioridad en otro lugar distinto; en CSS no existe el concepto de relaciones hermanas; CSS no es ni permite la incorporación de lenguajes de programación; CSS es incapaz de calcular cantidades o variables de almacenamiento; CSS no puede generar texto de forma independiente; CSS emplea un modelo simple de formateado de caja que no puede ser extendido para aplicaciones avanzadas de marcado de documentos, como el establecimiento de conjuntos de columnas; CSS no proporciona suficiente soporte para la internacionalización; etc. Para una más amplia información, véase Pankaj Kamthan. *XML and CSS: Structured Markup with Display Semantics* [documento HTML]. IRT.org, January 2, 2000. Disponible en <http://tech.irt.org/articles/js198/index.htm> (consultado el 8 de noviembre de 2000).

**XSLT** (*XSL Transformation*)<sup>134</sup> fue el primer lenguaje de los dos en desarrollarse, siendo recomendación del W3C desde noviembre de 1999. XSLT establece un conjunto de mecanismos para la manipulación y transformación de los documentos XML, bien sea para convertirlos a otros formatos (HTML, PDF, MSWord, etc.) o, bien, para realizar manipulaciones de DTDs o de documentos XML específicos y transformarlos en otras DTDs y documentos XML distintos. El caso más frecuente de utilización del lenguaje XSLT ha venido siendo la transformación de los documentos XML almacenados en un ordenador servidor al formato HTML cuando éstos son demandados a través de un cliente para su visualización en el espacio Web, dados los problemas de representación y visualización de los documentos XML en la mayoría de los *browsers* existentes en el mercado. La asignación de un formato de presentación de los documentos XML viene dado por el segundo lenguaje, el **XSL** (*Extensible Stylesheet Language*)<sup>135</sup>, también conocido como XSL-FO (*XSL Formatting Objects*), lenguaje aún en fase consolidación definitiva. Este lenguaje es el que verdaderamente se ocupa de establecer los mecanismos necesarios para el establecimiento de un formato o estilo de presentación del contenido existente en los documentos XML. Las propiedades de formateado de los objetos contenidos en los documentos XML son muy parecidas a las existentes para el lenguaje CSS pero a diferencia de éste, los documentos XSL son documentos redactados con la misma sintaxis que los documentos XML, existiendo, además, un número mayor de propiedades, lo que le convierte en un preciso y potente sistema de asignación de estilo para documentos XML. Su aprobación como Recomendación oficial del W3C no se ha producido hasta fechas muy recientes (octubre de 2001).

---

<sup>134</sup> World Wide Web Transformation. *XSL Transformation (XSLT) Version 1.0. W3C Recommendation* [documento HTML]. James Clark (ed.). W3C, November 16, 1999. Disponible en <http://www.w3.org/TR/xslt/> (consultado el 6 de febrero de 2001). En la actualidad se está trabajando en la siguiente versión de este lenguaje (versión 1.1), existiendo un *Working Draft* a fecha de 12 de diciembre de 2000, disponible en la dirección <http://www.w3.org/TR/xslt11/>

<sup>135</sup> World Wide Web Consortium. *Extensible Stylesheet Language (XSL) Version 1.0. W3C Recommendation* [documento HTML]. Sharon Adler... [et al.]. W3C, October 15, 2001. Disponible en <http://www.w3.org/TR/xsl/> (consultado el 21 de noviembre de 2001).

- **DOM (Document Object Model)**<sup>136</sup>: Para que un determinado programa informático sea capaz de procesar documentos XML resulta necesario que éste cuente con dos componentes básicos: un *parser* o analizador sintáctico del contenido del documento XML y la aplicación propiamente dicha encargada de tratar el contenido del documento XML a través del *parser* para su representación; esta aplicación puede ser desde un navegador o un editor hasta un sistema de almacenamiento y gestión en bases de datos. El *parser* y la aplicación deben compartir un mismo modelo de representación de los datos XML por lo que para ello se hace necesario un elemento intermedio entre ambos encargado de establecer dicha conexión entre ambos componentes: a este elemento informático se le conoce como API (*Application Program Interface*). Existen hoy en día dos tipos principales de APIs para XML: aquellas que se emplean una interconexión basada en objetos y aquellas que utilizan una interconexión basada en eventos. El primer caso se trata del DOM, ya visto con anterioridad en el apartado dedicado a la descripción del lenguaje HTML. Aunque el modelo DOM fue desarrollado originalmente por el W3C para unificar los modelos de representación de objetos HTML en los navegadores de Netscape Navigator 3 e Internet Explorer 3, esta recomendación se ha hecho extensible de igual modo para el tratamiento de los documentos XML. Así, el DOM define la estructura lógica de un documento XML representada por la simple jerarquía de sus elementos (o nodos, según DOM) creando un árbol de objetos en memoria. Este árbol jerárquico es una réplica exacta del árbol jerárquico de elementos existentes en el documento XML. Este modelo de API resulta perfecto para aquellas aplicaciones que su objetivo es la manipulación sencilla de los documentos XML a ser representados en, por ejemplo, un navegador o un editor. La interconexión basada en eventos resulta más sencilla que la anterior API. En este caso, el *parser* no necesita generar un árbol de objetos sino que, por el contrario, simplemente lee el contenido del fichero y genera una serie de eventos para localizar los elementos, atributos o el texto dentro del fichero. Así, estos eventos informan a la aplicación de que algo ha sucedido y ésta reacciona

---

<sup>136</sup> World Wide Web Consortium. "Document Object Model (DOM)..." *Op. cit.*,  
<http://www.w3.org/TR/REC-DOM-Level-1/> y <http://www.w3.org/TR/DOM-Level-2-Core/>

ante éste. La API más importante basada en eventos es **SAX** (*Simple API for XML*)<sup>137</sup> pero, a diferencia de DOM, ésta no está contemplada como un estándar oficial del W3C aunque la industria lo ha considerado como un estándar *de facto* dada su alta aceptación y empleo por parte de muchos usuarios. Esta API fue desarrollada a mediados de 1998 a través de las contribuciones aportadas por los miembros de la famosa lista de discusión electrónica XML-DEV<sup>138</sup> y definida, en un principio, para el lenguaje Java (en la actualidad existen versiones para otros lenguajes de programación como Python o Perl)<sup>139</sup>. El motivo de la gran aceptación que ha tenido SAX viene derivada de su alta eficacia en la optimización de los recursos empleados por la aplicación para procesar los documentos XML. Y esto es así, debido a que una API basada en eventos puede empezar a procesar el documento XML a medida que el analizador lo está leyendo (a diferencia de DOM que necesita leerlo por completo y representar el correspondiente árbol de objetos) lo que permite a la aplicación actuar de forma más rápida y consumir menos recursos del sistema. El principal problema de SAX viene dado por la imposibilidad de navegar a través del documento al prescindir de la representación de los elementos del mismo, por lo que resulta más conveniente para aquellas aplicaciones que mantienen sus propias estructuras de datos en un formato que no sea XML (por ejemplo, para la alimentación de bases de datos a través de documentos XML).

---

<sup>137</sup> Toda la información relativa al desarrollo de SAX, actualmente en su versión 2.0, se encuentra disponible en la dirección <http://www.megginson.com/SAX/index.html>

<sup>138</sup> Las contribuciones realizadas a esta lista de distribución por los más variados y prestigiosos investigadores en el desarrollo y avance del lenguaje XML pueden ser consultadas en la dirección <http://xml.org/archives/xml-dev/>

<sup>139</sup> Los principales impulsores en un principio de SAX dentro de esta lista fueron Peter Murray-Rust, Tim Bray y David Megginson. Este último fue el encargado de coordinar las posteriores contribuciones y el desarrollo evolutivo de este modelo de API en las diversas versiones que han ido surgiendo desde su lanzamiento. Para una mayor información, véase Aaron Skonnard. The “XML Files: XPath, XSLT, and others XML Specifications” [documento ASP]. *MSDN Magazine*, May 2000. Disponible en <http://msdn.microsoft.com/msdnmag/issues/0500/xml/xml0500.asp> (consultado el 15 de junio de 2000). De igual modo, en la anteriormente citada URL del sitio oficial de SAX se puede encontrar abundante información sobre la definición e historia de este desarrollo.

- **XML Schema Definition (XSD) Language**<sup>140</sup>: El modelo de esquema para documentos XML establecido por el W3C supone el establecimiento de un potente mecanismo para la formalización de nuevos vocabularios o lenguajes de definición estructural frente a las limitaciones existentes en el modelo de las DTDs, a las que sustituye. Dado que el mecanismo para la definición de nuevos lenguajes de marca en XML venía heredado del modo de proceder en SGML, este mecanismo estaba orientado, por tanto, casi con exclusividad al tratamiento de documentos de contenido textual y la definición de sus estructuras internas. Con la incorporación del entorno profesional de la informática, más acostumbrados al tratamiento de datos que de documentos, se hizo patente las grandes limitaciones del modelo original de DTD para dicho tratamiento. Así, desde febrero de 1999 el W3C comenzó a trabajar en la construcción de un modelo de definición de lenguajes de marcado descriptivo fuertemente influenciado por la modelización de datos llevadas a cabo en la construcción de las bases de datos relacionales, principalmente (proceso conocido como “esquema” de la base de datos, de ahí su nombre). Este nuevo modelo, entre otras muchas cosas, debería utilizar la misma sintaxis que los documentos XML (a diferencia de la DTD que ya sabemos que utiliza una sintaxis propia), así como ser capaz de tratar un gran número de tipos de datos (cadenas de texto, datos numéricos enteros y decimales, fechas, tiempos, datos binarios, etc.) o incluso poder ser definidos por el propio usuario, y ser capaz de soportar de forma adecuada importaciones y exportaciones de datos sobre cualquier sistema de gestión de bases de datos relacionales u orientadas a objetos. Resultado de este trabajo es la especificación del W3C conocida como XML Schema, establecida como Recomendación en mayo de 2001 y definida formalmente en tres partes:

---

<sup>140</sup> Toda la información oficial relativa al modelo de esquema XML del W3C se encuentra disponible en la dirección web del W3C en <http://www.w3.org/XML/Schema/>, documento creado y mantenido por Dan Connolly y Henry Thompson.

1. **XML Schema Part 0: Primer**<sup>141</sup>. Esta primera parte viene constituida por el texto introductorio de la especificación, en el cual se explica el significado y alcance de este modelo, las diferencias existentes frente a las DTDs y los mecanismos necesarios para la construcción de un nuevo esquema.
  2. **XML Schema Part 1: Structures**<sup>142</sup>. Esta segunda parte viene constituida por la propuesta de los diferentes métodos para la descripción de la estructura y la acotación de los contenidos de los documentos XML, así como la definición de una serie de reglas para la validación del esquema dentro de estos documentos.
  3. **XML Schema Part 2: Datatypes**<sup>143</sup>. La tercera y última parte viene constituida por la definición del conjunto de tipos de datos que pueden ser asociados a los elementos y atributos XML.
- **RDF (Resource Description Framework)**<sup>144</sup>: No se trata en realidad de un estándar asociado a XML y sí, más bien, de una aplicación o lenguaje específico de este metalenguaje (uno de los primeros en aparecer) pero de aplicación general a todo tipo de objetos electrónicos. RDF, basado en la idea lanzada por Berners-Lee de una “web semántica” (concepto que será tratado con mayor profundidad en próximos apartados de este capítulo), es, sin duda, el estándar actual más importante y con mayor

---

<sup>141</sup> World Wide Web Consortium. *XML Schema Part 0: Primer. W3C Recommendation* [documento HTML]. David C. Fallside (ed.). W3C, May 2, 2001. Disponible en <http://www.w3.org/TR/xmlschema-0/> (consultado el 21 de mayo de 2001).

<sup>142</sup> World Wide Web Consortium. *XML Schema Part 1: Structures. W3C Recommendation* [documento HTML]. Henry S. Thompson... [et al.] (eds.). W3C, May 2, 2001. Disponible en <http://www.w3.org/TR/xmlschema-1/> (consultado el 21 de mayo de 2001).

<sup>143</sup> World Wide Web Consortium. *XML Schema Part 2: Datatypes. W3C Recommendation* [documento HTML]. Paul V. Biron, Ashok Malhotra (eds.). W3C, May 2, 2001. Disponible en <http://www.w3.org/TR/xmlschema-2/> (consultado el 21 de mayo de 2001).

<sup>144</sup> Toda la información oficial sobre el estándar RDF (eventos y publicaciones oficiales, tutoriales, proyectos y aplicaciones, artículos y presentaciones, software, etc.) se encuentra disponible en la página del W3C coordinada por Ralph Swick y otros, “Semantic Web Activity: Resource Description Framework (RDF)” en la dirección <http://www.w3.org/RDF/>. Resulta también de especial la consulta del sitio de Dave Beckett, de la Universidad británica de Bristol, en la dirección <http://www.ilt.bris.ac.uk/discovery/rdf/resources/>

proyección para la descripción semántica de recursos en la web. Como se señala en la página oficial del RDF, este estándar integra a una gran variedad de actividades de aplicación en la Web que tienen a los metadatos como eje central de sus actividades, tales como los mapas de sitios web, descripción y clasificación de contenidos, definición de canales, recopilación de datos por parte de los buscadores de información en la web, desarrollo de colecciones en bibliotecas digitales y creación de documentos electrónicos de forma distribuida, utilizando en todos los casos XML como sintaxis de intercambio de datos entre aplicaciones. El modelo RDF se articula funcionalmente en tres grandes bloques: un modelo de datos, una sintaxis y un esquema. Tanto el modelo de datos como la sintaxis son formal y conjuntamente descritos en la Recomendación del W3C de 22 de febrero de 1999<sup>145</sup>. Como se señala en la propia especificación, ésta “establece un modelo para la representación de los metadatos RDF así como una sintaxis para la codificación y transferencia de los metadatos de modo que permita maximizar la interoperabilidad de forma independiente entre los servidores web y los clientes”. La sintaxis utilizada en RDF emplea, como decíamos, el lenguaje XML siendo, por tanto, ambos estándares complementarios entre sí: uno de los principales objetivos de RDF es la asignación de semánticas específicas normalizadas a los datos basados en XML. El modelo de esquema RDF, aún en fase de recomendación candidata<sup>146</sup>, no especifica un lenguaje o vocabulario concreto para elementos descriptivos sino, más bien, establece los mecanismos necesarios para la definición del tipo de contenido (datos) que han de llevar los elementos, las clases de recursos en los que pueden ser utilizados y poder así restringir las posibles combinaciones y relaciones entre estas clases y poder detectar de igual modo las violaciones a las restricciones establecidas. Esto es, un esquema RDF define un dominio específico de conocimiento semántico a través de la especificación

---

debido a la amplitud y rigurosidad científica de los múltiples enlaces seleccionados sobre diversos aspectos del estándar RDF.

<sup>145</sup> World Wide Web Consortium. *Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation* [documento HTML]. Ora Lassila, Ralph R. Swick (eds.). W3C, February 22, 1999. Disponible en <http://www.w3.org/TR/REC-rdf-syntax/> (consultado el 12 de septiembre de 2000).

de tipos de jerarquías y de definiciones de datos (basadas en el estándar ISO/IEC 11179 para la descripción de elementos de datos).

- **XML Query**<sup>147</sup>: Dada la estrecha relación que existe entre la estructuración de los contenidos en los documentos XML y el diseño y confección de bases de datos, resultaba obvio la necesidad de desarrollo de un potente lenguaje de interrogación capaz de expresar consultas complejas a través de todos los tipos posibles de datos XML (documentos estructurados y semi-estructurados, bases de datos relacionales y repositorios de objetos). A pesar de que en el entorno de XML existían otros estándares asociados capaces de producir consultas en documentos XML, como es el caso del empleo de XPath y XSLT para dicha función, éstos utilizaban una sintaxis poco adecuada y ciertamente alejada del tradicional lenguaje SQL (*Structured Query Language*), ampliamente utilizado en el mundo de las bases de datos relacionales. Debido a este vacío normativo, algunas de las principales compañías informáticas en este campo han venido desarrollando en estos últimos años sus propios lenguajes de interrogación para XML, como son, entre otros muchos, el caso de **XSQL**<sup>148</sup> (creado por David Beech de la compañía Oracle) o, más recientemente, el caso de **QUILT**<sup>149</sup> (desarrollado por Don Chamberlin (IBM), Dana Florescu (INRIA) y Jonathan Robie (Software AG)). El desarrollo XML Query del W3C recoge todas estas aportaciones y establece un lenguaje

---

<sup>146</sup> World Wide Web Consortium. *Resource Description Framework (RDF) Schema Specification 1.0. W3C Candidate Recommendation* [documento HTML]. Dan Brickley, R.V. Guha (eds.). W3C, March 27, 2000. Disponible en <http://www.w3.org/TR/2000/CR-rdf-schema-20000327/> (consultado el 12 de septiembre de 2000).

<sup>147</sup> Toda la información oficial relativa al lenguaje de interrogación XML se encuentra disponible en la dirección web del W3C en <http://www.w3.org/XML/Query/>, documento creado y mantenido en la actualidad por Massimo Marchiori.

<sup>148</sup> David Beech. *Position Paper on Query Languages for the Web* [documento HTML]. Oracle Corporation, 1998. Disponible en <http://www.w3.org/TandS/QL/QL98/pp/oracle.html> (consultado el 25 de febrero de 2001).

<sup>149</sup> Don Chamberlin, Jonathan Robie, Daniela Florescu. "Quilt: An XML Query Language for Heterogeneous Data Source". En: Dan Suciu, Gottfried Vossen (eds.). *The World Wide Web and Databases. International Workshop WebDB (3º. 2000. Dallas. Berlin [etc.]: Springer, 2001, pp. 1-25. En cualquier caso, para profundizar más sobre esta importante propuesta recomendamos la consulta de los documentos referenciados en la página web de este desarrollo en <http://www.almaden.ibm.com/cs/people/chamberlin/quilt.html>*

robusto capacitado para su aplicación en todos los tipos particulares de fuentes de datos XML existentes. Esta especificación, también denominada a veces como XQL (*XML Query Language*) o, simplemente, XQuery, se encuentra hoy en día definida más es su plano teórico que en su implementación práctica en el mundo real, dado que aún se encuentra en sus fases iniciales de discusión en el seno de la W3C. XML Query se desarrolla a través de cinco apartados: los requisitos necesarios, los posibles escenarios de uso, el modelo de datos, el álgebra y la sintaxis propiamente dicha.

1. **XML Query Requirements**<sup>150</sup>. Este primer documento especifica las metas y requisitos necesarios que este lenguaje ha de soportar.
2. **XML Query Use Cases**<sup>151</sup>. Continuación del anterior documento, en el que se amplían los contenidos anteriores y se establecen los escenarios posibles en los que ha de actuar este lenguaje de interrogación.
3. **XML Data Model**<sup>152</sup>. Este documento define el modelo de interrogación de datos XML, describiendo formalmente el conjunto de información disponible para realizar una consulta.
4. **The XML Query Algebra**<sup>153</sup>. Estas consultas o interrogaciones deben tener una representación abstracta definida mediante un álgebra específica, la cual describirá el núcleo de operaciones de este lenguaje.
5. **XQuery: A Query Language for XML**<sup>154</sup>. Es la sintaxis propiamente dicha de este lenguaje, expresada según lo definido por XML.

---

<sup>150</sup> World Wide Web Consortium. *XML Query Requirements. W3C Working Draft* [documento HTML]. Don Chamberlain... [et al.] (eds.). W3C, February 15, 2001. Disponible en <http://www.w3.org/TR/xmlquery-req/> (consultado el 17 de abril de 2001).

<sup>151</sup> World Wide Web Consortium. *XML Query Use Cases. W3C Working Draft* [documento HTML]. Don Chamberlain... [et al.] (eds.). W3C, February 15, 2001. Disponible en <http://www.w3.org/TR/xmlquery-use-cases/> (consultado el 17 de abril de 2001).

<sup>152</sup> World Wide Web Consortium. *XML Query Data Model. W3C Working Draft* [documento HTML]. Mary Fernandez, Jonathan Robie (eds.). W3C, February 15, 2001. Disponible en <http://www.w3.org/TR/xmlquery-datamodel/> (consultado el 17 de abril de 2001).

<sup>153</sup> World Wide Web Consortium. *The XML Query Algebra. W3C Working Draft* [documento HTML]. Peter Fankhauser... [et al.] (eds.). W3C, February 15, 2001. Disponible en <http://www.w3.org/TR/xmlquery-algebra/> (consultado el 17 de abril de 2001).

- **XForms**<sup>155</sup>: Esta aplicación XML, desarrollada para dar respuesta a las limitaciones de los actuales formularios HTML para la Web, se encuentra aún en fase de Working Draft<sup>156</sup>. Como se comenta en el propio documento de este grupo de trabajo, los actuales formularios web no diferencian entre el propósito y la presentación de los mismos siendo, por tanto, esta división entre lo que un formulario debe hacer y cómo ha de presentarse el primer punto de arranque de la especificación XForms. Además, el modelo planteado por este desarrollo pretende ser un dispositivo independiente y sencillo que sea capaz de trabajar con una amplia variedad de estándares XML así como en diferentes agentes de usuario (no sólo en navegadores web). Otro aspecto importante en este desarrollo está en la acertada idea de que la función principal de los formularios radica en la recolección de datos para su almacenamiento y/o intercambio con otras aplicaciones por lo que los nuevos formularios creados al abrigo de este estándar deberán reflejar perfectamente esa transacción estructurada de datos; es decir, un formulario definido mediante XForms debe dar origen o debe alimentarse de una instancia de datos XML, la cual está definida por una estructura previamente establecida mediante el modelo de DTD o esquema XML. Para llevar a cabo esta transacción entre el modelo formulado por XForms a través de la interfaz del agente de usuario y la instancia de datos XML es necesario establecer un canal de comunicación entre dichos elementos. Este hecho se obtiene a través del denominado *XForms Submit Protocol*. La relación de este estándar con otros estándares XML asociados es, lógicamente, muy estrecha, como es el caso de lo sucedido con, por citar algunos, el XHTML (para la presentación de dichos formularios en el tipo documental más común en la futura Web

---

<sup>154</sup> World Wide Web Consortium. *XQuery: A Query Language for XML. W3C Working Draft* [documento HTML]. Don Chamberlain... [et al.] (eds.). W3C, February 15, 2001. Disponible en <http://www.w3.org/TR/xquery/> (consultado el 17 de abril de 2001).

<sup>155</sup> Toda la información oficial relativa al lenguaje de interrogación XML se encuentra disponible en la dirección web del W3C en <http://www.w3.org/MarkUp/Forms/>, documento creado y mantenido en la actualidad por Thierry Michel.

<sup>156</sup> World Wide Web Consortium. *XForms 1.0. W3C Working Draft* [documento HTML]. Micah Dubinko... [et al.] (eds.). W3C, August 28, 2001. Disponible en <http://www.w3.org/TR/2001/WD-xforms-20010828/> (consultado el 25 de septiembre de 2001).

de Internet), con los esquemas XML (para la definición de los tipos de datos a incluir dentro de los campos del formulario) o los lenguajes CSS y XSL de hojas de estilo (para el establecimiento de las propiedades de presentación de estos formularios).



## **CAPÍTULO III.3**

### **APLICACIONES XML: XML ORIENTADO A DOCUMENTOS VS. XML ORIENTADO A DATOS**

### III.3.1. ORIENTACIÓN AL TEXTO Y ORIENTACIÓN A LOS DATOS EN XML

Definir XML resulta, como ya se ha expuesto en este capítulo, relativamente sencillo dado que no es más que una gramática para definir lenguajes de marcado de texto específicos; esto es, hacer posible que el usuario pueda crear etiquetas de marcado propias. Ahora bien, otra cosa bien distinta supone tener que ofrecer una respuesta sencilla acerca del uso y aplicaciones que sobre dicho metalenguaje se pueden establecer.

Antes incluso del establecimiento definitivo como recomendación oficial para la Web del W3C, Jon Bosak estableció en su famoso artículo *XML, Java, and the future of the Web*, citado con anterioridad en este capítulo, cuatro importantes áreas de aplicación para el nuevo metalenguaje de aplicación en la Web que estaba por llegar<sup>157</sup>:

1. Aplicaciones que requieren que el cliente web medie entre dos o más bases de datos heterogéneas.
2. Aplicaciones que tratan de distribuir una proporción significativa de procedimientos de carga desde el servidor web al cliente.
3. Aplicaciones que requieren que el cliente web presente de distintas formas unos mismos datos según sean los diversos tipos de usuarios a los que vayan destinados dichos datos.
4. Aplicaciones donde los agentes inteligentes de la Web traten de confeccionar información a la medida según sean las necesidades de los usuarios individuales.

Con el paso del tiempo y el empleo práctico de este estándar se fue comprobando que el campo mayor de aplicación se encontraba, sin duda, en la distribución e intercambio de datos en la Red. Este hecho fue constatado en un estudio estadístico realizado por J. Quin en octubre de 1999, y presentado al congreso anual de la ASIS (*American Society for*

---

<sup>157</sup> Jon Bosak. "XML, Java...". *Op. cit.*, <http://www.ibiblio.org/pub/sun-info/standards/xml/why/xmlapps.htm>

*Information Science*)<sup>158</sup>. Este autor demostraba que la mayor parte de las aplicaciones que del lenguaje XML se estaban realizando correspondían principalmente al sector del desarrollo de aplicaciones informáticas y a procesos de negocios en la red Internet (figura 1) siendo, por tanto, este tipo de organizaciones comerciales e industriales las que de forma mayoritaria estaban apostando por el desarrollo de vocabularios XML (figura 2).

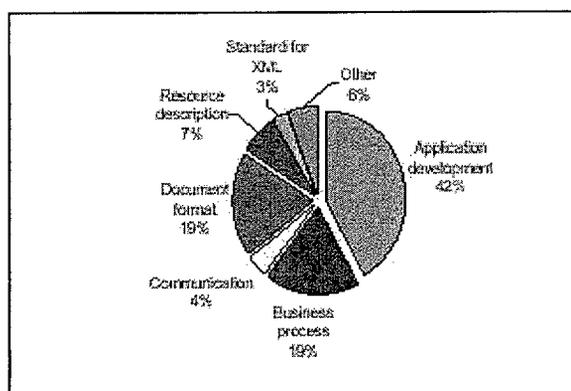


Figura III.2: Áreas de aplicación de XML.

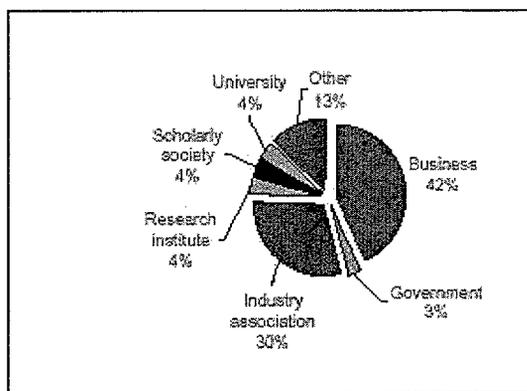


Figura III.3: Organizaciones involucradas en el desarrollo de aplicaciones XML.

Fuente: J. Quin. "Discipline- and industry- wide metadata schemas: semantics and Namespaces Control".

En la actualidad, resulta ciertamente difícil predecir hacia dónde se van a orientar las aplicaciones XML. Sin duda, las áreas de mayor pujanza seguirán siendo las mismas que se señalaban en el anterior estudio dado que es donde se encuentra el mayor nicho de mercado y, al fin y al cabo, XML no deja de ser un suculento negocio para muchas compañías informáticas.

<sup>158</sup> Jean Qin. "Discipline- and industry- wide metadata schemas: Semantics and Namespace Control". En: ASIS Annual Meeting, 1999. Citado En: Jean Qin. "Representation and Organization of Information in the Web Space: From MARC to XML" [documento PDF]. *Informing Science*, v. 3, n° 2, 2000, p. 86. Disponible en <http://inform.nu/Articles/Vol3/v3n2p83-88.pdf> (consultado el 4 de julio de 2000).

Finalmente, E. Dumbill señala en un artículo reciente que, sin duda, existe una enorme diversidad de aplicaciones XML pero señala las siguientes áreas como aquellas en las que XML se ha desplazado de forma más reciente; éstas son<sup>159</sup>:

- **Informática distribuida:** Los trabajos que se están llevando a cabo sobre el protocolo SOAP (*Simple Object Access Protocol*) y el propio protocolo XML están tomando un importante peso específico como formato de comunicación entre ordenadores a través de redes informáticas.
- **Configuración:** XML es hoy en día el formato más popular para la configuración sencilla de ficheros electrónicos.
- **Servicios de directorio:** Existen diversas iniciativas XML (DSML, UDDI) que pretenden adaptar el popular protocolo de servicios de directorio LDAP (*Lightweight Directory Access Protocol*) a este nuevo entorno de trabajo.
- **Almacenamiento:** Desarrollos como WebDAV permiten la mejora en el almacenamiento y gestión de los datos contenidos en sistemas de ficheros remotos a través del uso de protocolos XML específicos. Además, hoy en día todos los fabricantes de bases de datos ofrecen, en mayor o menor grado, el almacenamiento, gestión y recuperación de datos XML.
- **Formatos de presentación de páginas:** Desarrollos como XSL-FO o el SVG (*Scalable Vector Graphics*) están teniendo una amplia aceptación y su futuro es más que prometedor.
- **HTML:** XHTML, y de forma particular la modularización XHTML, está reinterpretando la manera habitual de construcción de páginas web.
- **Gestión del conocimiento:** En este campo tan prometedor y tan de moda en nuestros días entran en juego tecnologías como los mapas de materias y RDF para la construcción y asignación de valores semánticos a los documentos y a los datos electrónicos.

---

<sup>159</sup> Edd Dumbill. *The State of XML: Why Individuals Matter* [documento HTML]. XML.com, May 30, 2001. Disponible en <http://www.xml.com/lpt/a/2001/05/30/stateofxml.html> (consultado el 25 de junio de 2001).

Sintetizando todas estas aplicaciones, y para los fines de interés perseguidos en esta tesis doctoral, resulta pues necesario señalar que este metalenguaje ha sido orientado en estos últimos años hacia dos grandes tipos de aplicaciones, como así es señalado por B. Marchal<sup>160</sup>:

1. **Aplicaciones orientadas al documento** (*document applications*), donde la información tratada va dirigida casi con exclusividad para su consumo por parte los seres humanos. En este caso, XML se concentra en la definición de estructuras documentales con fines a su posterior publicación con independencia del medio de salida (monitor del ordenador, impresora, dispositivos parlantes, etc.) En este tipo de aplicaciones suele ser también frecuente utilizar a XML como formato de medio independiente para, con posterioridad y de forma automática, convertir estos documentos XML de contenido principalmente textual a otros formatos de publicación más comunes y extendidos, como HTML, PostScript, PDF, RTF, MSWord, etc.
2. **Aplicaciones orientadas a los datos** (*data applications*), donde la información tratada va dirigida principalmente para su consumo por parte de los programas informáticos. Aquí, el elemento fundamental radica en el tratamiento y manipulación de los datos contenidos para su gestión e intercambio entre sistemas de bases de datos. En el caso de ser necesaria la publicación de dichos datos en un espacio electrónico se hará igualmente uso de las técnicas derivadas del estándar XSLT para asignarles un marco visual adecuado en dicha publicación.

Esta doble orientación en la aplicación de XML viene ocasionada debido en gran medida a la naturaleza dual de la DTD en XML pues ésta, como se ha comentado, es tanto un mecanismo para generar nuevas gramáticas en los lenguajes de marcado de documentos electrónicos así como un mecanismo para el establecimiento de un esquema de datos con el

---

<sup>160</sup> B. Marchal. *Op. cit.*, p. 28.

que caracterizar un tipo documental. Esta naturaleza dual de la DTD será tratada con mayor detalle en el siguiente apartado de esta sección.

Resulta, por todo ello, fácil de entender que XML pueda ser utilizado para múltiples propósitos con finalidades distintas y diversas orientaciones, hecho éste que, como acertadamente señala M. Floyd, ha venido generando gran confusión y serios debates entre las diversas comunidades de usuarios según sea su campo de procedencia profesional o investigadora<sup>161</sup>.

Para entender de forma correcta esta doble orientación hacia el documento (el texto) y hacia los datos nada mejor que mostrar el ejemplo comparativo de dos instancias de documento XML que ofrece el propio W3C, donde se ve claramente cómo la información contenida, siendo en muchos casos la misma, es marcada de forma estructuralmente distinta<sup>162</sup>:

Orientación al documento	Orientación a los datos
<pre>&lt;memo importance="high" date="1999-03-23"&gt; &lt;from&gt;Paul v. Biron&lt;/from&gt; &lt;to&gt;Ashok Malhorta&lt;/to&gt; &lt;subject&gt;Latest draft&lt;/subject&gt; &lt;body&gt; We need to discuss the latest draft &lt;emph&gt;immediately&lt;/emph&gt;. Either email me at &lt;email&gt;mailto:paul.v.biron@kp.o rg&lt;/email&gt; or call &lt;phone&gt;555- 9876&lt;/phone&gt; &lt;/body&gt; &lt;/memo&gt;</pre>	<pre>&lt;invoice&gt; &lt;orderDate&gt;1999-01-21 &lt;/orderDate&gt; &lt;shipDate&gt;1999-01-25&lt;/shipDate&gt; &lt;billingAddress&gt; &lt;name&gt;Ashok Malhotra&lt;/name&gt; &lt;street&gt;123 Microsoft Ave.&lt;/street&gt; &lt;city&gt;Hawthorne&lt;/city&gt; &lt;state&gt;NY&lt;state&gt; &lt;zip&gt;10532-0000&lt;/zip&gt; &lt;/billingAddress&gt; &lt;voice&gt;555-1234&lt;/voice&gt; &lt;fax&gt;555-4321&lt;/fax&gt; &lt;/invoice&gt;</pre>

<sup>161</sup> Michael Floyd. *Creación de sitios Web con XML*. Madrid [etc.]: Prentice Hall, 2000, p. 8.

<sup>162</sup> Este ejemplo tan ilustrativo se encuentra en la introducción de la tercera parte de la Recomendación del XML Schema, ya citada con anterioridad y analizada con mayor profundidad en el siguiente apartado de esta sección. Véase World Wide Web Consortium. "XML Schema Part 2: Datatypes...". *Op. cit.*, <http://www.w3.org/TR/xmlschema-2/>

Veamos detenidamente algo más de cada una de estas dos principales orientaciones en el uso y aplicación del metalenguaje XML.

### A. MARCADO ORIENTADO AL TEXTO

El origen de XML viene derivado, como sabemos, por el influjo de múltiples corrientes de trabajo en la aplicación de las tecnologías de la computación al procesamiento de los textos electrónicos. Así, por un lado, toda la herencia recibida del metalenguaje SGML hace que XML pueda ser aplicado de forma similar a éste, donde el acento en el marcado del texto electrónico va orientado casi con exclusividad a la descripción de la estructura y la semántica de los documentos para su posterior gestión, tratamiento y difusión del mismo a través de la asignación de un formato de presentación y visualización de los mismos. Aquí el marcado XML hace posible que los programas informáticos representen o impriman los documentos electrónicos según un formato de presentación establecido a tal fin, así como que sea posible la localización, extracción y manipulación de la información contenida en dichos documentos. Como acertadamente señala R. Bourret, los documentos XML que ponen un especial acento en el texto contenido están caracterizados por una estructura irregular y, por tanto, flexible, donde los modelos de contenido de los elementos suelen combinar el texto contenido con otros subelementos marcados, estando la información y datos descritos poco o nada normalizada en el documento original (hecho éste que obliga al empleo asiduo de atributos para normalizar la información textual o los datos contenidos o, asimismo, utilizar otros mecanismos existentes en XML para ello)<sup>163</sup>.

Esta orientación en el uso del marcado del texto ha sido la que tradicionalmente se ha venido empleando desde los entornos profesionales del campo de la edición y publicación editorial y, en gran medida, también del ámbito de la biblioteconomía y la documentación (archivos, bibliotecas y centros de documentación). Se trata, pues, de la faceta más

---

<sup>163</sup> Ronald Bourret. *XML and Databases* [documento HTML]. Darmstadt : Technical University of Darmstadt, 1 September, 1999. Disponible en <http://www.informatik.tu-darmstadt.de/DVS1/staff/bourret/xml/XMLAndDatabases.htm> (consultado el 25 de noviembre de 1999).

“documental” del empleo del metalenguaje XML donde el documento, con contenido principalmente textual, es rigurosamente descrito según una secuencia de bloques estructurales y elementos informativos que le permiten ser diferenciado frente a otros documentos (una sentencia de un tribunal de justicia se diferencia perfectamente, por ejemplo, de un texto teatral debido a estas características).

Pero, además, XML, al igual que SGML, permite que cada pieza estructural del texto del documento electrónico sea descrito por medio de la abstracción lógica que representa a través de la asignación de un nombre característico al elemento de marcado. Además, como decíamos anteriormente, cada elemento del marcado del texto (desde el elemento raíz que define al documento en su globalidad hasta el último en el árbol de jerarquía) puede contener una serie de atributos que permitan describir la metainformación (o metadatos) relativa a la parte del texto marcado, información en forma de datos que no está contenida directamente en el texto marcado del documento. Este método, ampliamente utilizado para el tratamiento de los documentos SGML, aunque válido para ciertos propósitos no resulta el más idóneo para la descripción de los metadatos en los documentos XML, siendo recomendable el empleo de otras tecnologías XML específicas para tal fin, como es el caso de RDF, como se expondrá en posteriores apartados de este capítulo.

En cualquier caso, este modo de proceder convierte a este metalenguaje, y según palabras de T. Freter, en una poderosa alternativa a los sistemas tradicionales de gestión documental (basados, como ya se comentó en anteriores capítulos, en la redacción de fichas electrónicas con las descripciones que caracterizan a los documentos almacenados)<sup>164</sup>.

K. Williams y sus colaboradores establecen que cuando XML se orienta al marcado de los documentos electrónicos de contenido principalmente textual este marcado vendrá caracterizado por dos importantes premisas<sup>165</sup>:

---

<sup>164</sup> Todd Freter. *XML: Document and Information Management* [documento HTML]. Sun Microsystems, September 8, 1998. Disponible en <http://www.sun.com/980908/xml/> (consultado el 9 de diciembre de 1999).

- Si las marcas introducidas en el documento son eliminadas, el texto contenido seguiría teniendo por sí mismo básicamente el mismo significado fuera de la instancia de documento XML.
- El orden o secuencia de aparición de la información (de los elementos constitutivos) es un factor crítico para entender el significado del documento, esto es, existe una estrecha relación entre el orden de marcado del texto del documento para su posterior presentación física (la estructura visual o geométrica) y la estructuración lógica del documento en cuestión: la DTD reproducirá la estructuración física de los contenidos del documento.

A esto, y según nuestro parecer, habría que añadir al menos tres hechos diferenciales más:

- Las DTDs XML que se orientan a la descripción de documentos suelen incluir muchos elementos con modelos de contenido mixtos (PCDATA junto con otros subelementos). Así, en las instancias de estos documentos XML suele ser frecuente encontrar, por ejemplo, párrafos de texto en los que dentro de los mismos se incluyen elementos utilizados para el establecimiento de relaciones hipertextuales o para identificar instituciones, personas, etc. Los modelos de contenido mixto son difícilmente reproducibles mediante un modelo de esquema orientado a las bases de datos dado que en éstas cada campo de una tabla sólo se permite la inclusión de datos textuales u objetos multimedia.
- Los documentos de carácter textual suelen contener datos escasamente estructurados lo que implica, por un lado, que los datos susceptibles de ser recuperados (la metainformación del documento) no siempre se encuentren en una posición fija o determinada dentro del documento y, por otro lado, que dichos datos normalmente no han sido preparados de forma normalizada para dicha labor de almacenamiento,

---

<sup>165</sup> Kevin Williams. *Professional XML Databases*. Birmingham: Wrox, 2000, p. 12.

transmisión y recuperación de la información contenida (nombres y apellidos de personas, nombres de instituciones, fechas, lugares geográficos, materias, etc.). En estos casos, la normalización de esos datos se suele realizar de forma frecuente a través de la inclusión de atributos dentro de las etiquetas iniciales de marcado XML. Este hecho conlleva que en aquellas instituciones donde la redacción de los textos (y, en especial, de aquellos datos considerados den especial interés para la búsqueda y recuperación documental) no se realiza siguiendo unas estrictas pautas de redacción el texto marcado no tenga ninguna utilidad funcional real, teniendo, por tanto, que proceder a la normalización de dicha metainformación a través, como decíamos, de la inclusión de atributos o, en otros muchos casos, mediante el uso de otros mecanismos de definición normalizada de contenidos, como es el caso de RDF.

- Por último, los documentos XML con esta orientación no tienen razón de ser si no son presentados ante un determinado medio de salida (normalmente será ante un navegador Web) para su lectura por parte del usuario, destinatario final de este producto. Este hecho conlleva que la asociación directa de uno o varios estilos de presentación a estos documentos para la publicación de los mismos en el espacio electrónico de la Web, normalmente a través de los lenguajes específicos asociados a XML (CSS y XSL), tome en este tipo de orientación (al igual que ocurría con el lenguaje HTML) un especial significado e importancia. Este hecho influirá enormemente a la hora de establecer las estructuras documentales y los vocabularios a través del mecanismo correspondiente pues en estos casos las DTDs o esquemas XML deberán tener en cuenta la presentación final del documento.

## **B. MARCADO ORIENTADO A LOS DATOS**

Frente a esta orientación en la aplicación del metalenguaje existe otra derivada del tratamiento de datos por parte del mundo profesional de la informática computacional aplicada a la construcción de bases de datos; esto es, XML orientado a los “datos”. Como

acertadamente señalan Goldfarb y Prescod en su manual de XML, muchos profesionales de este ámbito de actuación se percataron de que algunos documentos electrónicos, aquellos que están fuertemente estructurados, tienen un gran parecido con las bases de datos: se establece un “idioma” riguroso para representar el contenido de los documentos; se preservan los datos abstractos frente a su formato de presentación; las estructuras que se dan dentro de estos documentos, así como las relaciones que se establecen entre los elementos de los mismos, son fácilmente trasladables a un modelo de esquema de base de datos; etc.<sup>166</sup>.

En efecto, gran parte del entusiasmo generado por XML entre los profesionales de la informática ha venido derivado de esta orientación en su aplicación, aspecto éste fundamental para entender la evolución sufrida por algunos de los lenguajes específicos así como ciertos mecanismos asociados a este metalenguaje hacia postulados más orientados al tratamiento de datos, caso de la traslación del mecanismo de la DTD para definir nuevos vocabularios hacia los denominados esquemas XML.

Desde esta concepción XML puede ser entendido como otro formato más para la definición de archivos de datos dado que, como manifiesta M. Floyd, “es posible tomar cualesquiera datos basados en texto y estructurarlos y definirlos como elementos XML a definir”<sup>167</sup>. Esto es posible gracias a que XML es, en esencia, un potente y flexible formato de datos capaz de representar una amplia variedad de información, se encuentre ésta inserta en cualesquiera modelo de almacenamiento (bien directamente en documentos electrónicos o bien abstraída en bases de datos). Esta enorme potencialidad de XML para la estructuración de los datos resulta, como señalan J. A. Senso y A. de la Rosa, clave para una gestión óptima de la documentación<sup>168</sup>.

En este otro contexto, y a diferencia de lo que sucedía en la anterior orientación, el destinatario principal de estos datos no será el usuario final sino que, más bien, irán

---

<sup>166</sup> Ch. F. Goldfarb, P. Prescod. *Op. cit.*, p. 10.

<sup>167</sup> M. Floyd. *Op. cit.*, p. 8.

orientados al consumo y transferencia entre aplicaciones informáticas. De igual modo, aquí no resulta fundamental el orden de los elementos dentro del documento XML pues los datos pueden ser representados de múltiples formas diferentes y con órdenes diversos sin que por ello afecte a su funcionalidad<sup>169</sup>. Por otro lado, como señala P. Bourret, cuando se habla de datos éstos suelen crearse con una estructura o esquema de datos rígido previamente definido así como con unos formatos de tipos de datos normalizados (nombres, fechas, materias, etc.) donde apenas o nunca se mezclan dentro de un contenido textual; esto es, aquí los datos están estrictamente estructurados y definidos de forma más o menos aislada, lo que facilita enormemente su almacenamiento en bases de datos y su posterior recuperación y transferencia<sup>170</sup>.

Por último, y al contrario de la orientación al documento textual, no resulta de especial interés, en un principio, la asignación de formatos de presentación de estos datos pues en la mayoría de los casos los datos contenidos en estos documentos no serán visibles, como decíamos anteriormente, por el usuario final (por ejemplo, en el intercambio de datos entre aplicaciones informáticas dedicadas al comercio electrónico). En el caso de ser necesaria la publicación de los mismos en el espacio electrónico de la Web suele ser común en este tipo de aplicaciones orientadas a datos que éstos sean transformados e insertados dentro de formatos de presentación “documental” adecuados, especialmente en formato HTML, a través de los potentes mecanismos de transformación que se incluyen en el estándar XSLT,

---

<sup>168</sup> José A. Senso, Antonio de la Rosa. “Especificaciones XML aplicadas a la Documentación”. En: Maria Eulalia Fuentes i Pujol (dir.). *Anuari de Biblioteconomia, Documentació e Informació: BIBLIODOC 1999*. Barcelona: Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya, 1999, p. 134.

<sup>169</sup> Piénsese, por ejemplo, en cualquier base de datos relacional que almacena datos sobre las ventas de un producto (serían unas facturas con las fechas de compra y de envío, datos del comprador, datos del producto, etc.) trasladado a un modelo XML. Las columnas (correspondientes a los subelementos o atributos del documento XML) dentro de cada una de las tablas requeridas (correspondientes a los elementos de primer nivel en el documento XML) pueden tener el orden que se desee dado que la importancia aquí radica en el establecimiento de relaciones entre unas tablas y otras a través de las columnas oportunas. La presentación física de dicha factura podrá hacerse de múltiples y variadas formas del informe final (tabular, en columnas, etc.) y con los datos presentados en el orden y forma que se desee. Para una mayor información sobre esta idea, véase K. Williams. *Op. cit.*, p. 12.

<sup>170</sup> P. Bourret. “XML and Databases”. *Op. cit.*, <http://www.informatik.tu-darmstadt.de/DVS1/staff/bourret/xml/XMLAndDatabases.htm>

evitándose así los actuales problemas que existen para la presentación y visualización directa de documentos XML por parte de los navegadores Web.

Siendo esto así, resulta fácil comprender que exista una estrecha relación entre XML y el modelo de las bases de datos relacionales y, en especial, las orientadas a objetos: en el primer caso, y según palabras de S. St. Laurent y E. Cerami, XML puede describir fácilmente tanto las estructuras como los contenidos de las tablas de éstas así como las relaciones que se dan entre dichas tablas. En el caso de las bases de datos orientadas a objetos estos autores señalan que el potencial es aún mayor para el almacenamiento y tratamiento de documentos XML dado que las estructuras que soportan este tipo de bases de datos son más similares a la naturaleza jerarquizada de los documentos XML<sup>171</sup>. Las principales aplicaciones que se han venido desarrollando dentro de esta concepción del XML han estado íntimamente relacionadas como el intercambio y transmisión de datos entre ordenadores para, por ejemplo, el comercio electrónico a través de la red Internet y aplicaciones derivadas, donde este intercambio de datos se hace de forma “invisibles” para el usuario final.

Otros autores, a los cuales nos adherimos, señalan, sin embargo, que esta estricta división entre aplicaciones orientadas a documentos y aplicaciones orientadas a datos no debería de ser tan tajante y estricta. En gran medida, esta significativa fractura ha sido generada de forma artificial, basada en un deficiente conocimiento del significado y alcance del metalenguaje XML, centrándose más en lo que las herramientas informáticas actuales pueden hacer con XML y no en lo que se es capaz de realizar a través del enorme potencial de éste para todo este tratamiento. Además de las opiniones vertidas al respecto por parte de importantes autores procedentes del ámbito de aplicación de SGML, como, por ejemplo, Goldfarb o Sperberg-McQueen, resultan más que reveladoras las palabras de A. Navarro, C. White y L. Burman al respecto de este debate al señalar que “en realidad todos

---

<sup>171</sup> Simon St. Laurent, Ethan Cerami. *Building XML Applications*. New York [etc.]: McGraw-Hill, 1999, pp. 33 y ss.

los documentos XML son en sí mismo datos y, a su vez, todos los mensajes de datos XML son documentos o fragmentos de documentos”<sup>172</sup>.

Así, si XML se orienta exclusivamente como “algo similar” al lenguaje HTML, esto es, como un formato de presentación de documentos electrónicos para la navegación en el espacio Web, se estará perdiendo más de la mitad de todo su potencial. De igual modo, si XML es concebido sólo como un lenguaje de descripción de datos para el intercambio de información entre aplicaciones, se estará teniendo una visión muy limitada de su capacidad.

En realidad, como señalábamos con anterioridad, son muchos los casos (por no decir en la mayoría de ellos) en los que ambas orientaciones confluyen, tal vez, interesándonos en una primera instancia sólo por la edición, validación y publicación directa del documento XML para, con posterioridad, extraer los datos más significativos del mismo y tratarlos como tales. De igual modo, el proceso puede funcionar en orden inverso, pero en ambos casos se ha de tener presente desde un principio esta doble orientación. Esta unión entre la tradicional visión de los documentos y los modelos actuales de bases de datos es lo que para muchos investigadores supone, como señala N. Haakonstad, una de las mayores fascinaciones en la aplicación de XML al entorno de la Web dado que frente a las limitaciones que se tenían con el lenguaje HTML ahora es posible construir aplicaciones “híbridas” que traspasen las fronteras de los documentos y de los datos<sup>173</sup>.

Esta particular visión de los documentos electrónicos tratados mediante las tecnologías XML para la construcción y difusión de bases de datos ofrece un gran número de ventajas frente al modelo tradicional antes señalado, entre las que se pueden destacar las siguientes<sup>174</sup>:

---

<sup>172</sup> Ann Navarro, Chuck White, Linda Burman. *Mastering XML*. San Francisco: Sybex, 1999, p. 754.

<sup>173</sup> Norma Haakonstad. “XML: What HTML Wanted to Be!” [documento HTML]. En: *XML 98 Conference* (1998. Chicago). GCA, 1998. Disponible en <http://www.infoloom.com/gcaconfs/WEB/chicago98/Haakonstad.HTM> (consultado el 22 de enero de 2001).

<sup>174</sup> Bonifacio Martín Galán, J. Tomás Nogales Flores. *Ideas básicas sobre el mercado de documentos jurisprudenciales del Tribunal Constitucional mediante tecnologías XML* [documentos DOC]. Getafe: Departamento de Biblioteconomía y Documentación, Universidad Carlos III de Madrid, 3 de junio de 2001 (informe interno remitido a D. Ignacio Borrajo, Letrado de los Servicios Informáticos del Tribunal Constitucional).

1. El texto del documento electrónico, al estar señalado con una serie de marcas estructurales y semánticas, no necesita ser abstraído nuevamente a una serie de una serie de tablas, campos y filas (en el modelo de base de datos relacional). Al realizarse la descripción documental en el mismo momento de la construcción del documento electrónico, el texto contenido en estos documentos puede ser indexado automáticamente de forma precisa. La incorporación al sistema de un motor de búsqueda documental proporcionará los mecanismos necesarios para la recuperación de la información según su contexto.
2. El número de elementos (o campos) por los que puede ser realizada la búsqueda documental se incrementa tanto como la institución desee pues normalmente, además de los campos habituales de consulta de las bases de datos tradicionales, se señalan otros de interés para la posterior recuperación de la información contenida de una forma precisa.
3. Las búsquedas pueden realizarse dentro de un contexto estructural determinado lo que posibilita una mayor precisión y exactitud de las mismas. Por ejemplo, sería posible localizar aquellos documentos electrónicos donde una determinada persona, institución o palabra cuando aparezca situada dentro un campo determinado contenido por otro mayor.

A este tipo de modelo de base de datos claramente orientado a los documentos se le conoce en el entorno de producción de las tecnologías XML como *native XML Database*.

Precisando aún más estas ideas, se puede añadir que en las aplicaciones que se dan en la actualidad en el entorno de creación multimedia, donde se entremezclan documentos y datos con contenidos de todo tipo (texto, gráficos, ficheros de audio, ficheros de vídeo, etc.) es donde el lenguaje XML puede desatar todo su potencial desde la óptica más cercana a nuestra profesión de documentalistas. Cada una de estas piezas del contenido documental existe como tal en formato electrónico y, por tanto, es posible marcarlas, indexarlas, y reutilizarlas en una amplia variedad de nuevos contextos con propósitos diferentes. Esta concepción de los documentos/datos electrónicos viene de la mano de la denominada

“Gestión de contenidos” (*content management systems*), sistema que aúna los beneficios que reportan los sistemas tradicionales de gestión documental, el empleo de bases de datos relacionales y las orientadas a objetos y la gestión de conocimiento en las organizaciones<sup>175</sup>.

Sin embargo, es verdad, que aún hoy en día resulta ciertamente complejo aunar ambos mundos dado que algunos de los mecanismos que nos encontramos en la actualidad en el entorno de desarrollo de XML, como es el caso concreto del modelo de DTD XML, resultan ciertamente limitados para llevar a cabo esta fusión de aplicaciones diversas. Afortunadamente, este problema está siendo solventado con la aparición nuevos estándares en el seno del W3C, como es el caso del “XML Schema”, ya introducido con anterioridad y comentado con mayor detenimiento en el siguiente punto de este capítulo.

Finalmente, en esta frontera tan difusa entre documentos y datos, es necesario destacar la puesta en escena de una aplicación intermedia del metalenguaje XML, a caballo entre ambos mundos: los metadatos o, dicho de otro modo, un tipo especial de datos que informan sobre los contenidos de los documentos, describiendo de forma normalizada aquella información que es requerida para una perfecta catalogación, organización y recuperación de los documentos electrónicos. Este tercer aspecto de aplicación se verá con mayor profundidad en próximos apartados de este capítulo.

Partiendo de esta doble orientación en la aplicación del metalenguaje XML, han sido múltiples y diversos los desarrollos prácticos que han venido surgiendo en el mundo real desde la aparición del metalenguaje XML en 1998. En muchos casos se ha tratado únicamente de la adaptación de estas aplicaciones o lenguajes definidos en origen bajo la DTD de SGML a la sintaxis propia de la DTD para XML, lo que ha conllevado una conversión relativamente sencilla. Sin embargo, y dado el espectacular recibimiento sufrido

---

<sup>175</sup> Véanse al respecto de este tema los siguientes artículos de interés: el documento electrónico *Putting XML to Work: Advantages of Content Management* [documento HTML]. Interleaf, 1998. Disponible en [http://support.blenks.com/visual\\_basic/XML/putting\\_xml\\_to\\_work\\_content\\_management.htm](http://support.blenks.com/visual_basic/XML/putting_xml_to_work_content_management.htm) (consultado el 9 de diciembre de 1999) y, de elaboración mucho más reciente, el documento electrónico de P.G. Bartlett. “XML’s Promise: Delivering Customized Information Everywhere” [documento HTML]. En:

hacia XML por parte de todo tipo de colectivos, instituciones e investigadores, han venido surgiendo a lo largo de estos años un sinnúmero de nuevas iniciativas y propuestas para la definición formal de todo tipo de documentos y datos en formato electrónico, encontrándonos, pues, aplicaciones que van desde lenguajes para la definición y representación física de complejas fórmulas moleculares en el campo de la química, tales como el *Chemical Markup Language* (CML) o la creación de partituras musicales (*Music Markup Language* (MusicML) hasta para cuestiones tan, en principio, peregrinas como la codificación de datos genealógicos (*Genealogical Data Markup Language* (GedML) o la representación de información teológica y el marcado de textos litúrgicos (el *Theological Markup Language* (ThML) y el *Liturgical Markup Language* (LitML), respectivamente<sup>176</sup>.

Resulta, por todo ello, imposible señalar aún mínimamente las características de las principales aplicaciones prácticas (o DTDs) que se han venido desarrollando en los diversos campos del saber humano, emplazando al lector interesado a la consulta de las fuentes de información más importantes, tanto impresas como electrónicas, existentes en este apartado<sup>177</sup>.

---

*XML Europe 2001 Conference* (2001. Berlin). Disponible en <http://www.gca.org/papers/xml/europe2001/papers/html/sid-01-4.html> (consultado el 7 de agosto de 2001).

<sup>176</sup> Algunos de estos ejemplos tan peculiares e interesantes, difícilmente referenciados en otros recursos de información, han sido extraídos de la obra Michael J. Young, *Aprende XML Ya*. Madrid [etc.]: McGraw-Hill Interamericana de España, 2000, pp. 15 y ss.

<sup>177</sup> Resulta siempre complejo recomendar unas fuentes por encima de otras dado que cada una de ellas siempre aporta algún aspecto no cubierto por las demás. Aún así, nos aventuramos a exponer aquí aquellos recursos que se han considerado más interesantes para una mayor profundidad en el conocimiento de las principales aplicaciones que se han venido desarrollando bajo el metalenguaje XML. Como fuente impresa cabe destacar, aunque ya algo obsoleta, la obra de Simon St. Laurent, Robert Biggar, *Inside XML DTDs*. New York [etc.]: McGraw-Hill, 1999, donde además de encontrarnos unos excelentes capítulos iniciales sobre el alcance, usos y formas de definición de DTDs, se describen ampliamente las principales aplicaciones científico-técnicas que han venido surgiendo desde la aparición de este metalenguaje. En el espacio electrónico de la WWW caben destacar varios recursos de gran interés. En primer lugar, el W3C viene auspiciando de forma directa el desarrollo de algunas de estas importantes aplicaciones, caso, por ejemplo, de la recomendación RDF (*Resource Description Framework*) para la asignación de una sintaxis y un esquema XML propios que de soporte a los lenguajes de descripción e intercambio de recursos o metadatos en la Web, o del MathML, para la definición de fórmulas matemáticas y contenido científico que ha de ser suministrado en la Web. Toda la información oficial del W3C a este respecto se encuentra accesible desde la página principal de información sobre XML, referenciada con anterioridad, y disponible en la dirección <http://www.w3.org/XML/>. Otro recurso de información fundamental lo constituye el sitio Web mantenido por Robin Cover, referenciado igualmente con anterioridad en esta tesis para la localización de aplicaciones SGML, bajo el nombre de "The XML Cover Pages" y accesible en la Web en la dirección

En cualquier caso, y dado los propósitos de esta tesis doctoral, aquellas aplicaciones XML que se han venido elaborando en el ámbito del tratamiento de la documentación electrónica de valor jurídico tendrán su correspondiente análisis en los apartados finales de este capítulo, así como en el desarrollo del modelo práctico propuesto en el siguiente capítulo de esta tesis.

---

<http://xml.coverpages.org/> Entre la abundante información que suministra este recurso cabe destacar el apartado de aplicaciones SGML/XML, dividido en cuatro apartados: aplicaciones generales, aplicaciones académicas, aplicaciones gubernamentales y de la industria y aplicaciones XML propuestas. Por último, señalar por su importancia el recurso "The XML Schema Site", creado y mantenido por James Tauber, James Rice y Daniel Krech, y disponible en <http://schema.net/> Orientado exclusivamente a recoger los desarrollos más significativos realizados mediante XML donde estas DTDs o lenguajes específicos son agrupados según la categoría de aplicación a la que se adscriben (comercio y finanzas, educación, de propósito general, lenguaje y representación del conocimiento, metadatos y archivo, multimedia, científico y técnico, software y las orientadas al entorno de Internet y la WWW).