ANEXO 2. XML

2.1 Lenguajes de Marcas

En los sistemas informáticos los lenguajes utilizados pueden ser divididos en dos tipos: los que se utilizan para crear aplicaciones conocidos como lenguajes de programación, y los que sirven para "marcar" documentos, que es a los cuales pertenece XML.

Para entender la evolución del XML hay que retroceder a la época en la que los sistemas informáticos eran todos propiedad de las empresas que los creaban (década de los sesenta). A finales de los sesenta, la empresa IBM encargó a Charles F. Goldfarb que diseñara un sistema estándar para la gestión y edición de documentos, ya que muchos sistemas de IBM no podían comunicarse entre sí debido al distinto sistema de descripción que utilizaban. Charles Goldfarb, auxiliado por Ed Mosher y Ray Lorie, creó un lenguaje único de marcado que permitiese entenderse con los diferentes documentos generados por distintos sistemas y plataformas, reuniendo en una misma etiqueta el formato y la descripción del contenido. A este desarrollo se le denominó **marcado generalizado**, donde se acuñó el término **Lenguaje de Marcas**.

La idea del marcado generalizado era que cada etiqueta sirviera tanto para describir el aspecto exterior del texto (el formato) como para indicar su contenido (el tipo de información o dato), por lo que se diseñó un sistema muy completo y general capaz de dar solución a cualquier tipo de documento. La solución utilizaba etiquetas de descripción de datos relacionadas con plantillas de estilos de formato, consiguiendo así los dos objetivos. A este "lenguaje" se le denominó GML, Lenguaje de Marcas Generalizado (**GML**, Generalizad Markup Lenguaje).

Hasta 1974 se continuo el desarrollo del GML, y entre 1978 y 1986, el propio Goldfarb coordinó el equipo técnico que desarrolló la norma internacional ISO 8879, que describe lo que desde entonces se conoce el estándar del Lenguaje Generalizado de Marcas SGML (SGML, Standard Generalizad Markup Lenguaje)

Conocer las marcas que utiliza cada programa de tratamiento de documentos hace posible diseñar filtros que permiten traspasar la información de unos formatos de marcas a otros sin perder el diseño. La forma que IBM creó para solventar este problema de incompatibilidad entre documentos se basaba en tratarlos con marcas accesibles desde descripciones universales TXT, es decir, basadas en código universal ASCII, haciendo así posible su tratamiento desde cualquier sistema y plataforma. Los lenguajes de marcas son sistemas complejos de descripción de información, normalmente documentos que si se ajustan al estándar SGML, se pueden controlar desde cualquier editor ASCII.

Desde que el GML cayó en manos de ISO y lo convirtió en un estándar oficial en los años ochenta (ISO 8879) denominándose SGML, esta norma de carácter general se aplica para diseñar lenguajes específicos de marcas, cuyos ejemplos más conocidos son el Lenguaje de Marcado con hipertexto HTML (**HTML**, Hypertext Markup Leguage) y el Formato de Texto Enriquecido RTF (**RTF**, Rich Text Format).

2.2 El Lenguaje HTML

Otro de los sucesos importantes en la historia de los orígenes del XML es la aparición del lenguaje HTML.

El lenguaje HTML es originariamente un subconjunto del SGML especializado en la descripción de documentos en pantalla a través de marcas (tags, etiquetas). La facilidad de uso y la particularidad de que no es propiedad de nadie, hizo al HTML el sistema idóneo para compartir información en Internet. La expansión de Internet le ha dado una posición de privilegio y ha hecho que la idea inicial se modifique considerablemente.

En principio, la intención de HTML era que las etiquetas fueran capaces de marcar la información de acuerdo con su significado, sin importar cómo se mostraban en la pantalla. Lo importante era el contenido y no la forma, es decir, era un lenguaje de marcas orientado a describir los contenidos, dejando a cada visualizador web (browser) la tarea de dar formato al documento según su criterio. Esto daba lugar a que una aplicación podía presentar un documento según la interpretación del visualizador empleado; por este motivo HTML empezó a incluir la posibilidad de controlar el formato del documento con descripciones particulares, como es el caso de los atributos de las fuentes de letras o las más completas hojas de estilo en cascada, más conocidas como CSS (**CSS**, Cascading Style Sheets).

Por diversos motivos, los creadores de los navegadores fueron añadiendo más etiquetas HTML dirigidas a controlar la presentación, y los usuarios las utilizaron para que sus documentos estuviesen perfectamente formateados, sin permitir diferencias importantes entre visualizadores distintos, por lo que HTML pasó a ser un lenguaje de marcas más dirigido al control de la presentación; además, para facilitar la vida de los usuarios, los analizadores sintácticos de las marcas HTML que incluyen los navegadores permitieron saltarse algunas normas, el resultado es que HTML ya no es un lenguaje que sigue las normas estrictas de SGML.

Dado que HTML dejó de servir para su función inicial, el consorcio World Wide Web (W3C) se puso en la tarea de diseñar un nuevo subconjunto del SGML que sirviera para describir contenidos de documentos, al que se ha denominado XML (Extensible Markup Language) publicando las especificaciones de la versión 1.0 en el año de 1998.

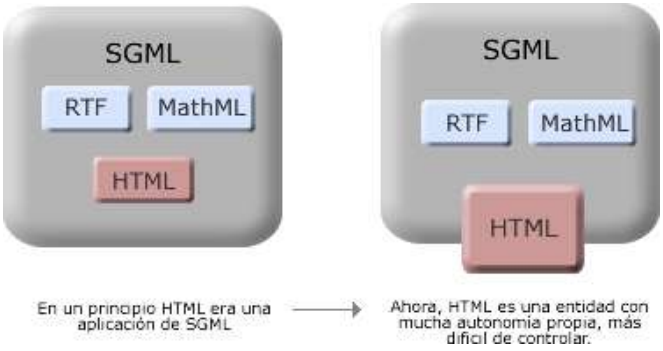

Figura 2-1. HTML y SGML

2.3 El Lenguaje XML

En teoría, HTML es una aplicación del XML especializada en la presentación de documentos para la web, mientras que XML es un subconjunto de SGML más fácil de tratar que este último y especializado en la gestión de todo tipo de información para la web. En la práctica, HTML tiene una parte dentro de XML (y en consecuencia queda dentro de SGML) y otra parte fuera de XML (y en algunas ocasiones también fuera de SGML).

Para reconducir esta situación, el consorcio W3C ha publicado reglas expresas para distinguir el HTML que sigue rigurosamente las normas, denominándolo XHTML (XHTML, Extensible Hypertext Markup Language), que explicado en forma muy breve, no es más que una reformulación de HTML dentro de las normas de XML.

Figura 2-2. XML

El objetivo de XML no se limita a ser una especificación que solucione las salidas de tono del HTML, sino que ha sido también pensado para servir de código unificado para los muy diversos aparatos y dispositivos que cada vez más se van conectando a Internet, como los aparatos de televisión, los hornos, los teléfono, etc.; además de proporcionar un estándar que permite la integración e interoperabilidad de varios tipos de sistemas.

XML, al igual que el SGML, es lo que se conoce como un "metalenguaje", es decir, un lenguaje (de marcas) capaz de generar otros lenguajes (de marcas). Este motivo es el que le hace ser el "padre" del XHML y de otros lenguajes específicos para determinadas tareas, como el WML (WML, Wíreles Markup Language) para teléfonos móviles, el VML (Vector Markup Language) para diseño gráfico vectorial, el SMIL (Sinchronized Multimedia Integration Language) para la presentaciones multimedia, y otros similares que poco a poco van apareciendo para solucionar diferentes problemas y ser aplicados en diferentes sectores.

Una aproximación rigurosa a la definición de XML es : El lenguaje XML es un lenguaje de marcas, basado en SGML, capaz de describir cualquier tipo de información en forma personalizada, aunque también es un metalenguaje capaz de describir lenguajes de marcas adecuadas para aplicaciones concretas. No obstante, XML también puede ser visto como un conjunto de normas que permiten tratar información muy diversa desde muchos puntos de vista y sistemas diferentes, siendo el propio diseñador el encargado de decidir el proceso más adecuado a cada caso y, en consecuencia, XML es un sistema complejo de descripción de información libre y rigurosa.

2.3.1 Terminología XML

2.3.1.1 DTD y Schemas

Es muy normal que en un proceso de tratamiento de información XML participen varias personas especializadas en diversos sectores o fases del proyecto o pertenecientes a diferentes departamentos o empresas, en cuyo caso no deben existir ambigüedades o enfoques distintos, ya que cualquier pequeña diferencia puede ocasionar errores o defectos en la gestión de la información. Por este motivo se han creado normas concretas para que la descripción de los datos quede especificada, de forma tal, que cualquier participante en el proceso sepa a que atenerse y se vea obligado a tratar la información de la misma forma que los restantes colaboradores, eliminando así la posibilidad de producir fallos.

Estas normas reguladoras de la estructura de los documentos XML se conoce como Definiciones de Tipos de Documentos o DTD (**DTD**, Document Type Definitions). Una definición de tipo de documento es una lista normas que describe con exactitud la composición que debe mantener la estructura de datos de cualquier documento XML.

Un documento que cumple con las normas básicas gramaticales de un documento XML se le denomina **Bien formado,** y cuando además cumple con las reglas estructurales de una determinada DTD se dice que el documento es **Válido**.

No es obligatorio que exista una DTD para que un documento XML tenga una aplicación práctica, pero si sí existe, todos los documentos, códigos, plantillas, etc., relacionados con dicho documento deben tener en cuenta la DTD para que todo funcione correctamente.

Para facilitar la declaración de las particularidades que deben seguir los documentos XML, Microsoft desarrolló una normativa que permitía hacer lo mismo que una DTD, pero con las diferencias fundamentales de que utiliza un lenguaje derivado del XML (más fácil de entender y manejar) y se apoya en otra especificación que permite describir los tipos de datos (Data Types). Estas normas han sido asumidas y reformadas por el W3C, denominándose **Esquemas** XML (XML Schemas).

2.3.1.2 DOM y SAX

Gran parte del éxito de XML se debe a la proliferación de librerías para procesarlo; y esta proliferación ha sido promovida por el consorcio W3C que en conjunto con la especificación XML ha creado especificaciones sobre los métodos de procesarlo. Existen dos especificaciones o métodos para procesar código XML. En el primer método, DOM (**DOM**, Document Object Model), se lee el documento completo y se identifica su estructura jerárquica. El segundo método, SAX (**SAX**, Standard API for XML), consiste en ir identificando las marcas a medida que se va leyendo el documento. El segundo método es obviamente más rápido y consume menos recursos, pero tiene la desventaja de que cada vez que aparece una marca se debe decidir que hacer con ella, y no se puede regresar para atrás en el documento.

Es bueno aclarar que tanto DOM como SAX, son sólo especificaciones. A las implementaciones de estas normas se les denomina comúnmente "Parsers". los "parsers" DOM se les denomina "tree based parsers", y a los de SAX "event driven parser".

2.3.1.3 XSL

XSL es el lenguaje de páginas de estilo que ha sido desarrollado por el consorcio W3C, para dar formato a los documentos XML, se llama XSL, que es el acrónimo de Extensible Style-sheet Language. Una página de estilo XSL permite modificar un documento XML, produciendo un resultado que puede estar en uno de varios formatos diferentes incluyendo el propio XML, HTML o un fichero de texto.

XSL se compone de un lenguaje de **transformación** y otro lenguaje de **formato**. El lenguaje de transformación XSLT (**XSLT**, XSL Transformations) recoge un documento XML bien formado o válido y lo transforma en un documento XML alternativo, y el lenguaje de formato XSL-FO (**XSL-FO**, XSL Formatting Objects) describe cómo debe ser visualizado el resultado de la transformación. De acuerdo con esto, un documento XML que esté vinculado a un documento XSL realiza en primer lugar una transformación de su contenido y, más tarde, se muestra de acuerdo al formato descrito para el documento resultante.