LOS LENGUAJES DE MARCADO APLICADOS A LOS REGISTROS BIBLIOGRÁFICOS. XML MARC DTD; XML MARC SCHEMA.

I. INTRODUCCIÓN A LOS LENGUAJES DE MARCADO

Los **lenguajes de marcado**, también denominados lenguajes de marcas, provienen y toman su nombre de la práctica tradicional e histórica de marcar los manuscritos con instrucciones de impresión en los márgenes. Esta tarea, tan habitual en la industria editorial desde la aparición de la imprenta, fue dando lugar a un grupo de **marcas estandarizadas**, cuya esencia ha sido trasladada hoy al mundo de la informática.

Actualmente, se distinguen **dos tipos fundamentales de lenguajes de marcado**, si bien en la práctica pueden combinarse:

- **LENGUAJES DE MARCADO DE PROCEDIMIENTO**, orientados hacia la **presentación del texto**. Sus símbolos o marcas indican la clase de operaciones tipográficas que deben ser aplicadas a cada uno de los elementos del documento electrónico, para dar **formato** al texto. Su misión es, pues, configurar la **apariencia física** de éste (fuente, estilo, tamaño de letra, etc.), tanto en pantalla como impreso.

- **LENGUAJES DE MARCADO DESCRIPTIVO**, diseñados para identificar las piezas o **bloques estructurales** que componen el texto. Sus marcas determinan la **estructura lógica** del documento electrónico y/o la **descripción de su contenido**, no ya su tipografía ni el formato que presentará cada fragmento en su posterior visualización.

Así pues, en el contexto automatizado actual, los lenguajes de marcado permiten **codificar documentos** intercalando, junto al texto, etiquetas o marcas provistas de información adicional sobre su estructura o presentación. En realidad, más que de lenguajes, podría hablarse de **metalenguajes o conjuntos de reglas** que se encargan de establecer y definir la forma digital de los documentos, bien para controlar su procesamiento, bien para representar su significado.

ORIGEN Y EVOLUCIÓN

A pesar de existir proyectos anteriores, como el **GenCode**, la iniciativa que sentó las bases de los lenguajes de marcas actuales partió en los **años 60** del investigador de IBM **Charles F. Goldfarb**, considerado hoy padre de este tipo de lenguajes por su participación en el desarrollo del *Generalized Markup Language* o **GML**.

La generalización de los lenguajes de marcas: SGML (años 80)

El gran éxito del lenguaje GML propició su extensión e incluso adopción por parte del gobierno de los EE.UU., hecho que acrecentó la necesidad de estandarizarlo. Así, tras un largo proceso, en **1986**, se convirtió en **norma ISO** bajo el nombre de **SGML** (*Standard Generalized Markup Language*).

ISO 8879:1986

A pesar de su falta de precisión y dificultad, SGML fue la **piedra angular** o **referente definitivo** de los lenguajes de marcado modernos, ya que, además de proveer una sintaxis para la inclusión de marcas en los textos, introdujo por primera vez una sintaxis para especificar qué etiquetas estaban permitidas y dónde. Éste sería el **punto de partida de dos conceptos clave**, abordados más adelante: **DTD** (*Document Type Definition*) y **Schema**.

La popularización: HTML (años 90)

El **Lenguaje de Marcas de Hipertexto** o **HTML** (*HyperText Markup Language*) fue definido en **1990**, a partir de la sintaxis de SGML, por **Sir Tim Berners-Lee**, creador de la WWW. En la actualidad, constituye la tecnología predominante en la construcción de páginas web, dada su simplicidad para estructurar textos y su capacidad para establecer enlaces con otros archivos. El **Consorcio World Wide Web (W3C)** regula las recomendaciones y versiones normalizadas de este lenguaje, aceptado como **norma ISO** desde el año **2000**.

ISO 15445:2000

La estructura HTML clave se basa en el uso de **ELEMENTOS**, compuestos mayoritariamente por **etiquetas**, **atributos** y **contenido**:

- **Etiquetas**. Están representadas por corchetes angulares (< >) con una instrucción en su interior. Cuando los elementos presentan una etiqueta inicial, texto y una etiqueta final, se denominan llenos. Cuando no presentan contenido ni necesitan cerrarse mediante etiqueta alguna se conocen como elementos vacíos.

elemento con argumento

- **Atributos**. Se incluyen en la etiqueta inicial del elemento pertinente, justo detrás de la instrucción (estructura nombre**=**"valor"), y recogen la variable a aplicar.

- **Contenido**. Es el texto propiamente dicho, consignado entre las etiquetas y posteriormente reconocido como información en la visualización HTML.

Todo documento HTML presenta **dos zonas claramente diferenciadas**: cabecera (***HEAD***) y cuerpo (***BODY***).

- La **cabecera** alberga el título del documento que aparecerá en la parte superior de la ventana del navegador así como otros datos no explícitos a posteriori, pero de interés para el servidor o para los buscadores web.

- El **cuerpo** contiene la información que el usuario verá en su navegador.

Puesto que HTML fue diseñado, en origen, para intercambiar información en entornos académicos, sus etiquetas están eminentemente pensadas para la **organización lógica del contenido** (titulo, párrafo, etc.) y no tanto para su presentación. Por eso, el W3C tuvo que idear las denominadas **hojas de estilo** (*Style Sheets*).

La madurez: XML (años 2000)

La respuesta a las deficiencias surgidas en el entorno HTML fue la aparición, en **1998**, del **lenguaje XML** (e***X****tensible* ***M****arkup* ***L****anguage*), desarrollado y evolucionado por el **W3C**.

Se trata de un lenguaje **sencillo**, que propicia la compatibilidad entre sistemas, permite incluir enlaces multidireccionales y cuenta a su alrededor con tecnologías complementarias, como el **Lenguaje de Hojas de Estilo Extensible** (**XSL***, E****x****tensible* ***S****tyle* ***L****anguage*), promovido también por el W3C para la presentación de documentos XML.

Desde el punto de vista estructural, XML guarda ciertas similitudes con HTML, ya que sus documentos cuentan con dos partes claramente delimitadas, **PRÓLOGO** y **CUERPO**, y sus componentes presentan denominaciones parecidas. Si bien el **prólogo** es opcional, el **cuerpo** es completamente obligatorio, pues recoge el contenido en sí del documento, dividido en texto y marcado, con **ELEMENTOS** (llenos o vacíos).

Como en HTML, los componentes clave del marcado son las **etiquetas** y sus **atributos**, aunque también cobran especial importancia los **comentarios** o las denominadas **entidades predefinidas** y **secciones CDATA**, ambas necesarias para representar caracteres propios de las marcas, que, en vez de procesarse como tal, deben figurar en la visualización posterior. Los datos consignados entre las marcas constituyen el **contenido**, perfectamente legible y sin codificación alguna.

Aunque se denomina "**extensible**", porque no limita el número de marcas o etiquetas posibles, es un lenguaje especialmente **estricto** en cuanto a lo que está permitido y lo que no. Así, todo documento XML debe cumplir **dos condiciones**:

- Estar **bien formado** (*well-formed*), es decir, guiarse por lo especificado en la recomendación W3C.

- Ser **válido** (*fully validated*), es decir, respetar las restricciones establecidas por una definición externa, **DTD** o **XML Schema**. Pero, ¿qué son exactamente ambos conceptos?

II. INTRODUCCIÓN A LOS CONCEPTOS *DTD* Y *SCHEMA*

DTD (Definición del Tipo de Documento)

Una DTD define los tipos de elementos, atributos y entidades permitidos así como sus posibles combinaciones. Las primeras DTD fueron desarrolladas en los laboratorios de **IBM**, hacia **1978**, cuando aún se experimentaba con el futuro lenguaje SGML. Actualmente, las más comunes son las usadas para entornos HTML y, sobre todo, para XML.

Su función básica consiste en **describir la estructura y sintaxis de los datos**, proporcionando un **formato común** que dé consistencia a todos los documentos regidos por la misma DTD. Aunque ésta puede incluirse dentro del propio documento, normalmente se almacena en un fichero aparte vinculado.

Las **limitaciones** y la **rigidez** de las DTD favorecieron la aparición de otras herramientas de descripción estructural, alternativas y más completas, como los denominados **Schemas**.

SCHEMAS

Un Schema es **similar a una DTD** en el sentido de que define qué elementos están permitidos, cómo deben organizarse y qué tipo de atributos pueden albergar, pero añade **VENTAJAS** como:

- El uso de la **sintaxis de XML**.
- Una **mayor especificación** del tipo de datos.
- Su **extensibilidad**.

El resultado, esto es, **XML Schema**, es un **lenguaje de esquema** utilizado para describir, con total precisión, la estructura y restricciones de contenido de los documentos XML, más allá de las normas sintácticas impuestas por el propio XML. Fue desarrollado por el **W3C** y alcanzó el nivel de **recomendación** en mayo de **2001**.

III. LENGUAJES DE MARCADO APLICADOS A LOS REGISTROS BIBLIOGRÁFICOS MARC

Los **formatos de marcas** constituyen desde hace tiempo una poderosa alternativa, o complemento, a los sistemas tradicionales de codificación de datos para describir recursos electrónicos. Los proyectos más importantes en este ámbito, liderados desde los años 90 por la *Library of Congress*, siguen, en la actualidad, **dos líneas fundamentales de trabajo** relacionadas con los lenguajes de marcado:

- Diseño de **modelos de descripción alternativos a MARC**, como el Dublin Core (DC).
- **Adaptación de los modelos tradicionales MARC** (norma ISO 2709) a los nuevos formatos de Internet.

Para muchos especialistas, el formato MARC debe seguir primando frente a otros sistemas de metadatos propuestos, como el DC, ya que:

- Sus **más de 30 años como soporte** de la comunidad bibliotecaria y de la industria de software especializado lo avalan. La mayoría de opciones restantes se encuentran aún en fase de desarrollo, no pudiéndose garantizar su permanencia en un futuro.
- Posee **mayor capacidad expresiva**, además de una **semántica** claramente definida, aceptada y adaptada a las distintas realidades nacionales.
- Da cabida a una **amplia gama de registros** (datos bibliográficos, autoridades, fondos y localizaciones, etc.).
- Se ha convertido en la **base de los SIGB actuales**.

Por éstas y otras razones, se han realizado diversos **esfuerzos para ligar MARC e Internet**, empezando por la inclusión del **campo 856** (*electronic location and access*) o la **adaptación de MARC a SGML** durante los años 90. Para ello, se crearon **dos DTD** capaces de convertir registros MARC a SGML y viceversa, sin pérdida de información, que pronto migrarían a XML para adecuarse a las nuevas necesidades tecnológicas.

MARC DTD

Oficina de Desarrollo de Redes y Normas MARC

En **2002**, la *Network Development and MARC Standards Office* (NDMSO) de la **LC** publicó un **esquema XML**, yendo más allá de un mero mecanismo de conversión y facilitando la representación de registros MARC en formato XML, para eliminar complejidades innecesarias y evitar que MARC quedara relegado frente a otras propuestas, en el marco de la biblioteca electrónica.

MARC SCHEMA

IV. MARC XML DTD

Las MARC XML DTD, derivadas de las mencionadas DTD para SGML iniciales, **se dividen en**:

1. **XML DTD** para **registros bibliográficos**, **fondos** e **información a la comunidad**.

2. **XML DTD** para registros de **autoridad** y **clasificación**.

Surgieron en **2001**, con los objetivos relatados, de acuerdo con los siguientes **PRINCIPIOS**:

1º **Generalidad**, por su independencia respecto a aplicaciones específicas basadas en MARC.

2º **Reversibilidad**, por su capacidad para convertir datos de una estructura a otra y poder volver al formato original, sin pérdida de contenido intelectual o de elementos semánticos esenciales.

3º **Flexibilidad**, ya que ofrece diversas posibilidades, en vez de imponerlas directamente.

4º **Amigabilidad** para el usuario (del inglés *user-friendly*), gracias a sus jerarquías lógicas.

5º **Relación con TEI** (*Text Encoding Initiative*), otra iniciativa sólida para el intercambio de información textual y para la descripción normalizada de documentos electrónicos.

V. MARC XML SCHEMA

Con este nuevo **avance** de la NDMSO de la **LC** se ha logrado una conversión más simple y flexible de los registros MARC, como norma ISO 2709, al lenguaje XML, adaptado a la WWW. En concreto, la utilización de la tecnología **XML Schema**, para definir la estructura de los registros MARC actuales, reporta **VENTAJAS** como:

- **Soporta todos los datos codificados con MARC**, independientemente de la variante utilizada, **y los transforma** para que sean perfectamente legibles en un entorno web.

- La **conversión** de MARC a XML y la **recomposición** de XML a MARC se realiza sin pérdida de información alguna.

- Posee una **arquitectura extensible** para describir recursos originales en sintaxis XML, permitiendo realizar modificaciones individuales o por lotes.

- Posibilita la **transformación** de MARC XML **a otros formatos de metadatos**.

La **gran mejora** de esta tecnología con respecto a las XML DTD radica en que, cuando se traduce un registro MARC con XML Schema, se mantienen todos y cada uno de los campos, indicadores y subcampos MARC, tal cual. Las DTD, por el contrario, obligan a redactar de nuevo toda esa información en formato XML, lo cual acaba por crear una DTD muy grande y con muchas más líneas.

VI. CONCLUSIÓN

En definitiva, los lenguajes de marcado son herramientas indispensables para la estructuración y presentación de los recursos bibliográficos en línea. Hoy, además, se postulan como elementos clave en el diseño de la **Web Semántica**, aquella que no sólo permite acceder a la información, sino también definir su significado, para facilitar su procesamiento automático.

XML, a través de sus **DTD y Schemas**, ha hecho de **MARC** un poderoso estándar descriptivo, también en el ámbito electrónico. No obstante, no faltan las **voces disonantes** que advierten de su complejidad, lentitud y coste ante grandes cantidades documentales. Por eso, han surgido otras alternativas, como el **DC** o el lenguaje **RDF**, que ofrecen mayor rapidez descriptiva, aunque en menor detalle.

PUNTOS TRATADOS EN EL TEMA

LOS LENGUAJES DE MARCADO APLICADOS A LOS REGISTROS BIBLIOGRÁFICOS. XML MARC DTD; XML MARC SCHEMA.

I. INTRODUCCIÓN A LOS LENGUAJES DE MARCADO

ORIGEN Y EVOLUCIÓN

La generalización de los lenguajes de marcas: SGML (años 80)

La popularización: HTML (años 90)

La madurez: XML (años 2000)

II. INTRODUCCIÓN A LOS CONCEPTOS DTD Y SCHEMA

DTD (Definición del Tipo de Documento)

SCHEMAS

III. LENGUAJES DE MARCADO APLICADOS A LOS REGISTROS BIBLIOGRÁFICOS MARC

IV. MARC XML DTD

V. MARC XML SCHEMA

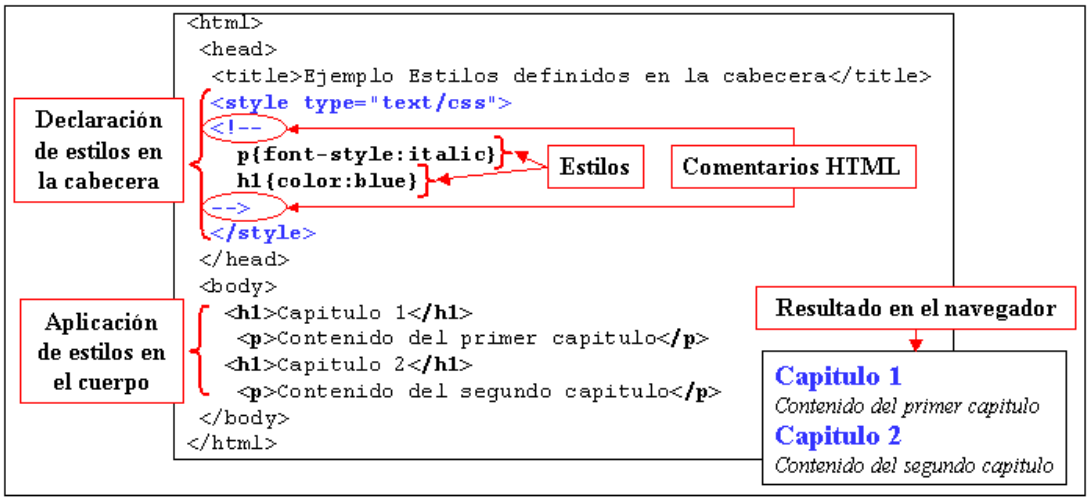
VI. CONCLUSIÓN

ANEXO

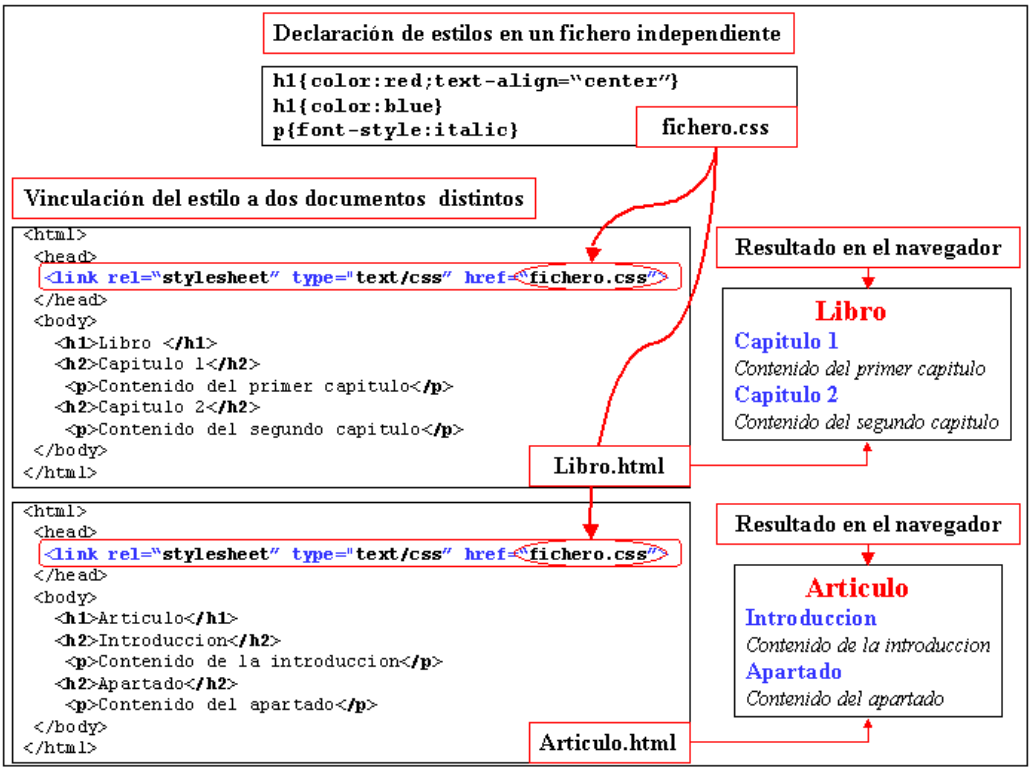
Reglas de la hoja de estilo aplicadas a un elemento en HTML:

```
<p  style="color:blue">Mediante  el  atributo  style  se  puede  alterar  el
aspecto del elemento al que se le aplica.</p>

RESULTADO (aplicado al elemento párrafo):

Mediante el atributo style se puede alterar el aspecto del elemento al que
se le aplica.
```

Reglas de la hoja de estilo aplicadas a la cabecera de un documento en HTML:



Reglas de la hoja de estilo vinculadas (a uno o varios documentos) como fichero externo:



* El atributo **rel** indica el enlace a una hoja de estilo. El atributo **type** indica el tipo de hoja de estilo (en este caso texto) y el atributo **href** hace referencia a la dirección URL donde está ubicado el **fichero.css**.