UD. Qué es el XML

- Qué es el XML
- Conceptos y vocabulario
- Documentos bien formados
- Documentos válidos
- Otras recomendaciones XML

---

1. Qué es el XML

El XML (eXtensible Markup Language = Lenguaje de Marcas Extensible) no es un lenguaje de marcas, sino un metalenguaje, es decir, el XML define las reglas generales que debe cumplir un lenguaje de marcas y la manera de definir un lenguaje de marcas.

El XML fue creado por el W3C a finales de los 90. El W3C se creó en 1994 para tutelar el crecimiento y organización de la web. Su primer trabajo fue normalizar el HTML, el lenguaje de marcas con el que se escriben las páginas web. Al crecer el uso de la web, crecieron las presiones para ampliar el HTML. El W3C decidió que la solución no era ampliar el HTML, sino crear unas reglas para que cualquiera pudiera crear lenguajes de marcas adecuados a sus necesidades, pero manteniendo unas estructuras y sintaxis comunes que permitieran compatibilizarlos y tratarlos con las mismas herramientas. Ese conjunto de reglas es el XML, cuya primera versión se publicó en 1998.

Curiosamente, el HTML no cumple las normas del XML ya que el creador del HTML, Tim Berners-Lee se basó en el SGML, otro conjunto de reglas para la creación de lenguajes de marcas creado en los años 80 más complejo que el XML, para definir el HTML. El W3C aprobó en el año 2000 el XHTML, una versión del HTML que sí que cumple las reglas del XML. El W3C pretendió sin éxito que el HTML dejara de utilizarse y sólo se utilizara XHTML. Al no conseguirlo, el W3C decidió retomar el desarrollo del HTML (incluyendo en él una versión XHTML).

Por su parte, el éxito del XML ha sido enorme y cada vez es más utilizado como sistema de intercambio y almacenamiento de información. El W3C ha desarrollado alrededor del XML numerosas tecnologías para sacar provecho del XML.

---

2. Conceptos y vocabulario

Documento XML

Un documento XML es un documento de texto plano (sin formato).

Procesador XML (XML processor) y aplicación (application)

Cuando una aplicación necesita leer un documento XML, la aplicación recurre a un procesador XML. El procesador XML (o analizador XML, en inglés XML parser) es el que lee el documento, analiza el contenido y le pasa la información en un formato estructurado a la aplicación. La recomendación XML especifica lo que debe hacer el procesador, pero no entra en lo que hace después la aplicación con esa información.

Caracteres (characters)

Los documentos XML pueden estar codificados en distintos juegos de caracteres (iso-8859-1,

utf-8, etc).

Marcas (mark-up) y contenido (content)

El texto que contiene un documento XML se divide en marcas y contenido. Las marcas pueden ser de dos tipos: etiquetas o referencias a entidades. Todo lo que no son marcas es contenido.

Etiquetas (tags)

Una etiqueta es una marca que empieza con el caracter "<" y termina con ">". Existen tres tipos de etiquetas:

- las etiquetas de apertura (start-tag). Por ejemplo:

  ```
  <apartado>
  ```

- las etiquetas de cierre (end-tag), que empiezan por "/". Por ejemplo:

  ```
  </apartado>
  ```

- las etiquetas vacías (empty tag), que terminan por "/". Por ejemplo:

  ```
  <salto-de-linea />
  ```

Referencias a entidades

Una entidad consiste en un nombre y su valor (son similares a las constantes en los lenguajes de programación). Las entidades se definen mediante la etiqueta ENTITY, por ejemplo:

```
<!ENTITY yo "Miguel de Cervantes">
```

Una referencia a una entidad empieza con el caracter "&", sigue con el nombre de la entidad y termina con ";". Al abrir el documento XML el procesador sustituye la referencia a la entidad por su valor. Por ejemplo, la etiqueta:

```
<autor>&yo;</autor>
```

el procesador XML la convertiría en:

```
<autor>Miguel de Cervantes</autor>
```

Existen varias entidades predefinidas, necesarias para poder utilizar los caracteres que delimitan las marcas o las cadenas de texto:

Referencia a entidad	Carácter
&lt;	<
&gt;	>
&amp;	&
&apos;	'
&quot;	"

Elementos (elements)

Un elemento es un componente lógico de un documento que o bien comienza por una etiqueta

de apertura y termina por la etiqueta de cierre correspondiente o que consiste en una única etiqueta vacía. El contenido de un elemento es todo lo que se encuentra entre las etiquetas de apertura y cierre, incluso si estos son también elementos en cuyo caso se llaman elementos hijos.

Atributos (attributes)

Un atributo es un componente de las etiquetas que consiste en una pareja nombre (name) / valor (value). Se puede encontrar en las etiquetas de apertura o en las etiquetas vacías, pero no en las de cierre. En una etiqueta no puede haber dos atributos con el mismo nombre. La sintaxis es siempre nombreAtributo="valorAtributo". . Por ejemplo:

```
<autor nombre="Miguel" apellidos="de Cervantes" />
```

Instrucciones de procesamiento (PI, processing instruction)

Una instrucción de procesamiento en una etiqueta que empieza por "<?" y acaba por "?>" y que contiene instrucciones dirigidas a las aplicaciones que leen el documento. Pueden aparecer en cualquier lugar del documento. Por ejemplo:

```
<?xml-stylesheet type="text/xsl" href="estilo.xsl" ?>
```

Declaración XML (XML declaration)

La declaración XML es una etiqueta que comienza por "<?xml " y termina por "?>" y que proporciona información sobre el propio documento XML. Aunque no es obligatoria es conveniente que aparezca y debe aparecer siempre al principio del documento. No es una instrucción de procesamiento, pero tiene la misma sintaxis (empieza por <? y acaba por ?>). Por ejemplo:

```
<?xml version="1.0" encoding="iso-8859-1"?>

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

Es importante que el juego de caracteres que aparece en la declaración sea el juego de caracteres en que realmente está guardado el documento, porque si no el procesador XML puede tener problemas leyendo el documento.

Definición de tipo de documento (DTD, Document Type Definition)

Una DTD es un documento que define la estructura de un documento XML: los elementos, atributos, entidades, notaciones, etc, que pueden aparecer, el orden y el número de veces que pueden aparecer, cuáles pueden ser hijos de cuáles, etc. El procesador XML la utiliza para verificar si un documento es válido, es decir, si el documento cumple las reglas del DTD.

Declaración de tipo de documento (DOCTYPE, Document type declaration)

Una declaración de tipo de documento es una etiqueta que comienza por "<!DOCTYPE" y acaba por ">" y que indica la DTD que debe utilizar el procesador XML para validar el documento. La DTD puede estar incluida en el propio documento o ser un documento externo. Por ejemplo:

```
 <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
```

Comentarios (comments)

Un comentario es una etiqueta que comienza por "<!--" y acaba por "-->". Los comentarios no pueden estar dentro de otras marcas y no pueden contener los caracteres "--". Dentro de un comentario las entidades de carácter no se reconocen, es decir, sólo se pueden utilizar los caracteres del juego de caracteres del documento. Por ejemplo:

```
<!-- Esto es un comentario -->
```

Secciones CDATA (CDATA section)

Una sección CDATA es una etiqueta que comienza por "<![CDATA[" y termina por "]]>" y cuyo contenido el procesador XML no interpreta como marcas sino como texto. Es decir que si aparecen los caracteres especiales (< & " ') en una sección CDATA, el procesador XML no interpreta que empieza un marca sino lo considera un carácter más. Se suele utilizar en documentos en los que aparecen muchas veces esos caracteres especiales para no tener que estar utilizando las referencias a entidades (< & &quot ') que hacen el texto bastante incómodo de leer .

Documentos bien formados

Un documento XML debe estar bien formado, es decir debe cumplir las reglas de sintaxis de la recomendación XML. Para que un documento esté bien formado, al menos debe cumplir los siguientes puntos:

- El documento contiene únicamente caracteres Unicode válidos.
- Hay un elemento raíz que contiene al resto de elementos.
- Los nombres de los elementos y de sus atributos no contienen espacios.
- El primer carácter de un nombre de elemento o de atributo puede ser una letra, dos puntos (:) o subrayado (_).
- El resto de carácteres pueden ser también números, guiones (-) o puntos (.).
- Los caracteres "<" y "&" sólo se utilizan como comienzo de marcas.
- Las etiquetas de apertura, de cierre y vacías están correctamente anidadas (no se solapan) y no falta ni sobra ninguna etiqueta de apertura o cierre.
- Las etiquetas de cierre coinciden con las de apertura (incluso en el uso de mayúsculas y minúsculas).
- Las etiquetas de cierre no contienen atributos.
- Ninguna etiqueta tiene dos atributos con el mismo nombre.
- Todos los atributos tienen algún valor.
- Los valores de los atributos están entre comillas (dobles).
- No existen referencias en los valores de los atributos.

Si un documento XML no está bien formado, no es un documento XML. Los procesadores XML deben rechazar cualquier documento que contenga errores.

Documentos válidos

Un documento XML bien formado puede ser válido. Para ser válido, un documento XML debe:

- incluir una referencia a una gramática
- incluir únicamente elementos y atributos definidos en la gramática
- cumplir las reglas gramaticales definidas en la gramática

Existen varios formas de definir una gramática para documentos XML, las más empleadas son :

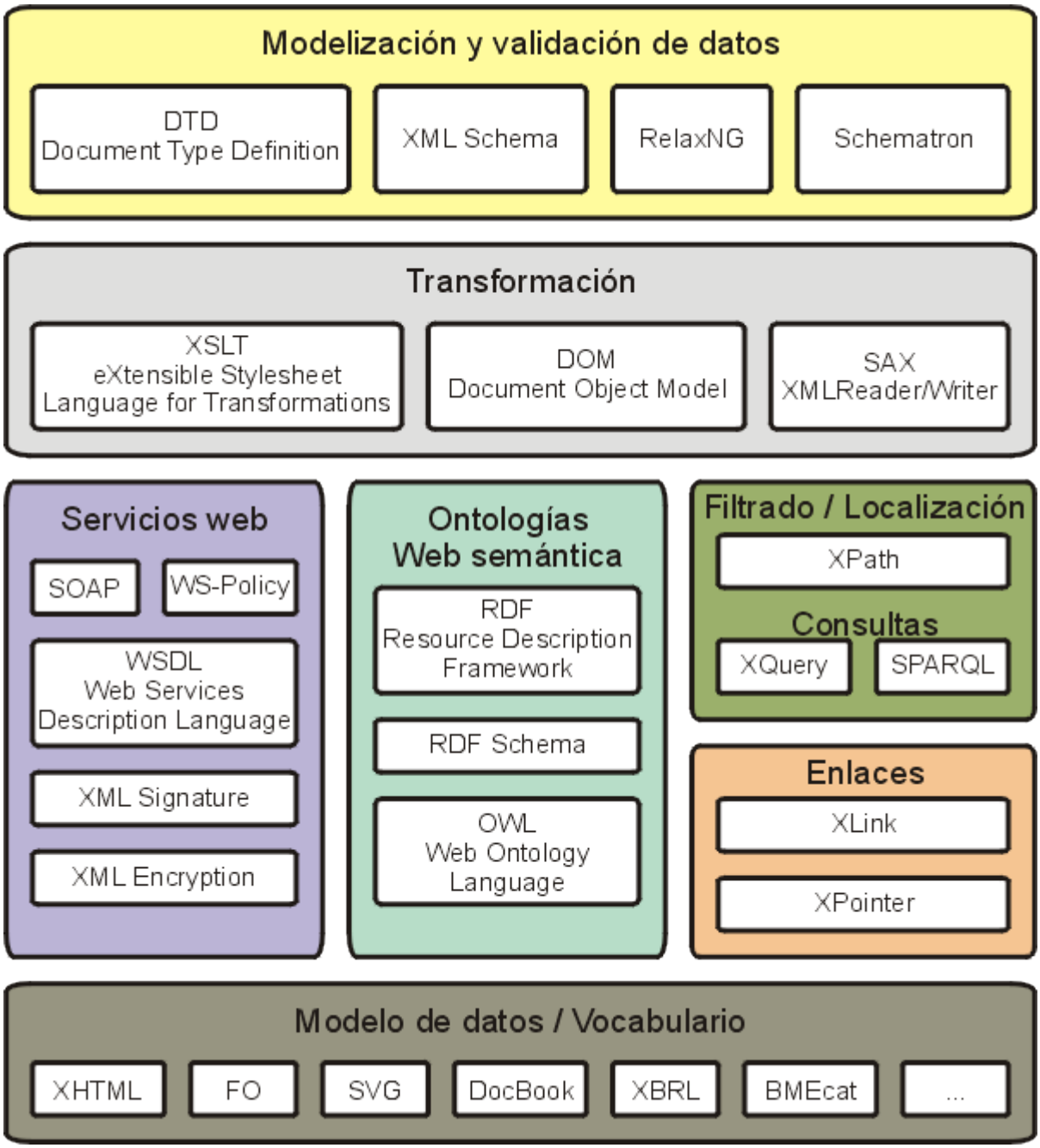
- DTD (Document Type Definition = Definición de Tipo de Documento). Es el modelo más

antiguo, heredado del SGML.

- XML Schema. Es un modelo creado por el W3C como sucesor de las DTDs.
- Relax NG. Es un modelo creado por OASIS, más sencillo que XML Schema.

Otras recomendaciones XML

El W3C y otras organizaciones de normalización han publicado numerosas recomendaciones relacionadas con XML. El cuadro siguiente cita algunas de ellas agrupándolas por temas:



Las más empleadas son las siguientes:

XML Namespaces (Espacios de nombres XML)
: Define los mecanismos para permitir que en un documento se utilicen elementos y atributos

de diferentes vocabularios, sin tener que preocuparse de que algunos nombres coincidan.

XML Base
: Define el atributo xml:base, que puede utilizarse como base para resolver las referencias a URI relativas en un elemento XML

XML Infoset
: Describe un modelo de datos abstracto para documentos XML a partir de elementos de información. Se utiliza en las especificaciones de lenguajes XML, para describir restricciones en el lenguaje.

xml:id
: Define el atributo id

XPath
: Define las expresiones XPath que sirven para identificar los componentes de un documento XML y facilitar su acceso a los programas que procesan documentos XML.

XSLT
: Lenguaje de transformación de documentos XML a otros formatos (XML o no XML)

XSL Formatting Objects (XSL-FO)
: Lenguaje de marcas para formatear documentos XML que se usa, por ejemplo, para generar PDFs.

XQuery
: Lenguaje de consulta orientado a XML, que permite acceder, manipular y devolver fragmentos de documentos XML.

XML Signature
: Define la sintaxis y las reglas de procesamiento para crear firmas digitales en documentos XML.

XML Encryption
: Define la sintaxis y las reglas de procesamiento para encriptar documentos XML.

Otras recomendaciones del W3C relacionadas con el XML no han tenido mucho éxito, como XInclude, XLink y XPointer.