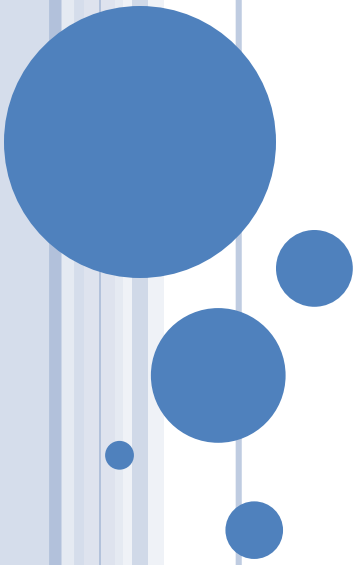


XML eXtensible Markup Language

Lenguajes de Marcas y Sistemas de Gestión de la Información

Profesora: Inés Menéndez

XML

- XML (*eXtensible Markup Language*, lenguaje de marcado extensible) es un lenguaje de marcado de propósito general
- Fue creado por el W3C a finales de los años 90. Se puede consultar este documento: http://www.w3.org/TR/1998/REC-xml-19980210
- Actúa como metalenguaje para definir otros lenguajes
  - Vocabularios XML
- Es una simplificación de SGML orientada a la web
- En XML no existen marcas predefinidas como en HTML
  - Debe definirse un conjunto de marcas apropiado para la información que se quiere describir en cada caso
- Al igual que en HTML, las marcas se escriben entre los signos '<' y '>'
  - Documento = contenido + marcas

Objetivos de XML

- XML se debe poder utilizar directamente en Internet
- XML debe admitir una gran variedad de aplicaciones
- XML debe ser compatible con SGML
- Debe ser fácil crear programas que procesen documentos XML
- El número de funcionalidades opcionales de XML deberá mantenerse en un mínimo absoluto, preferiblemente cero
- Los documentos XML deberán ser inteligibles para los humanos y razonablemente claros
- El diseño de XML deberá prepararse rápidamente
- El diseño de XML deberá ser formal y conciso
- Los documentos XML deberán ser fáciles de generar
- La concisión en las marcas XML tiene una importancia mínima

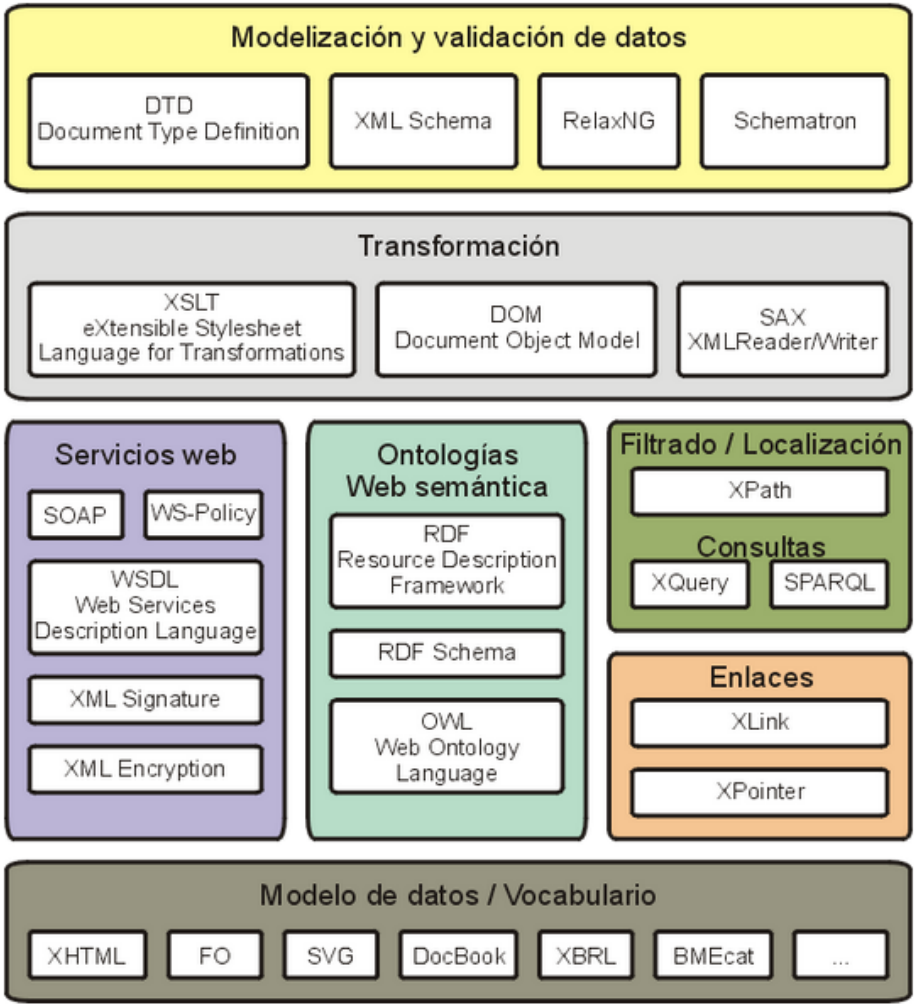
Lo que NO es XML

- No es un lenguaje de programación, de manera que no existen compiladores de XML que generen ejecutables a partir de un documento XML.
- No es un protocolo de comunicación, así que no enviará datos por nosotros a través de Internet. (Los protocolos de comunicación como HTTP o FTP sí pueden enviar documentos con formato XML)
- No es un sistema gestor de bases de datos, aunque una base de datos puede contener datos de tipo XML y existan bases de datos nativas XML, que lo que almacenan son documentos XML.
- No es propietario, esto es, no pertenece a ninguna compañía.

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:



Ejemplo de documento XML

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!- Este es un ejemplo de documento XML -->
<!DOCTYPE anuncio SYSTEM "anuncioDTD.dtd">
```

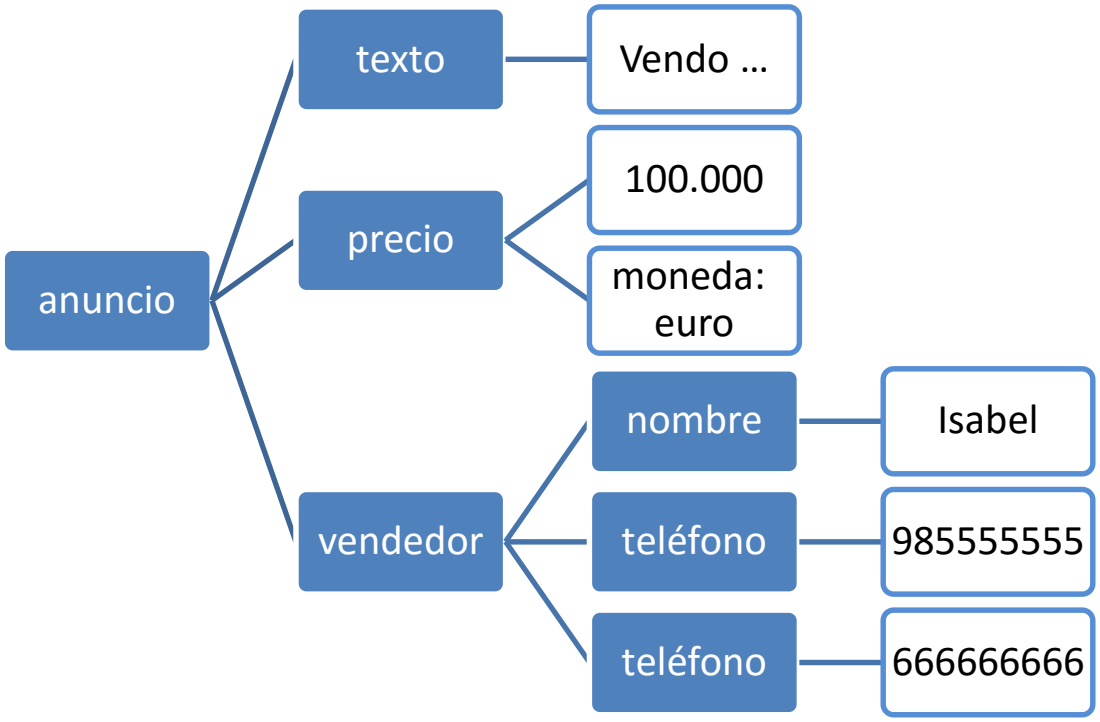
Prólogo

```
<anuncio>
    <texto>vendo piso de 3 dormitorios en Gijón</texto>
    <precio moneda="euro">100000</precio>
    <vendedor>
        <nombre>Isabel</nombre>
        <telefono>985555555</telefono>
        <telefono>666666666</telefono>
    </vendedor>
</anuncio>
```

Cuerpo

Modelo de datos de un documento XML

- El documento se organiza jerárquicamente en forma de árbol
  - Los nodos internos son los elementos
  - Las hojas son los contenidos



Componentes de un documento XML

- Prólogo
  - Declaración XML
  - Instrucciones de procesamiento
- Contenido - cuerpo
  - Elementos
  - Atributos
  - PCDATA
  - CDATA
  - Espacios en blanco
- Comentarios
- Notaciones
- Entidades

Partes de un documento XML

- Un documento XML tiene dos partes
  - Prólogo
    - Es una zona de declaraciones
    - Es opcional, aunque es recomendado
- Elemento documento o elemento raíz
  - Es el elemento principal del documento, que contiene a todos los demás
  - Tiene que haber exactamente uno

Prólogo

- Declaración XML
  - información de la versión de XML utilizada
  - conjunto de caracteres utilizado para codificar la información
  - ¿documento aislado? (standalone="yes|no")

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="yes"?>
```

- Declaración del tipo de documento (opcional)
  - Define el tipo y estructura del documento (sintaxis)

```
<!DOCTYPE anuncio SYSTEM "anuncioDTD.dtd">
```

Prólogo

- Instrucciones de procesamiento (opcional)
  - Se marcan entre ‘<?’ y ‘?>’
  - Proporcionan información que el procesador pasará a la aplicación XML
    - Dependiente del procesador del documento
    - P.e. Vincular una hoja de estilo a un documento

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

Elementos

- Componentes básicos de un documento XML
- El identificador de los elementos debe comenzar con una letra y puede contener caracteres de subrayado y de dos puntos pero no espacios en blanco
- Los elementos se delimitan mediante etiquetas formadas por el identificador
- Todos los elementos deben tener las etiquetas de inicio y de fin
  - <identElemento>contenido del elemento</identElemento>
- Los elementos vacíos tienen una forma abreviada
  - <identificadorElementoVacio/> (<A/> equivale a <A></A>)

Elementos

- El contenido de un elemento puede estar compuesto de:
  - Otros elementos
  - Datos formados por caracteres
  - Referencias a otras entidades
  - Mixto (mezcla de los anteriores)
- Opcionalmente los elementos pueden contener atributos
  - Metainformación acerca de los datos de los elementos

Contenido o datos formados por caracteres

- El contenido o datos formados por caracteres, es cualquier texto que no son marcas
  - Contenido textual de los elementos
  - Valores de los atributos
  - Un literal de cadena (“dato” o ‘dato’)
- Los caracteres menor que (<) y el ampersand (&) no pueden pertenecer al contenido
  - Debido a que tienen un significado especial en el marcado
  - Se pueden utilizar empleando secuencias de escape **<** y **&**

Atributos

- Un elemento puede contener atributos que proporcionen información adicional sobre dicho elemento
- Los atributos no se consideran parte del contenido de un documento
- Los atributos se utilizan para asociar pares nombre-valor a los elementos
  - Los valores de los atributos están formados por cadenas de caracteres
  - El valor debe escribirse entre comillas, simples o dobles
  - El orden en el que aparecen es indiferente
  - No puede haber más de un atributo con el mismo nombre para un mismo elemento
- La especificación de los atributos debe aparecer sólo dentro de las etiquetas de inicio o de las etiquetas de elementos vacíos

Elementos vs Atributos

- En ocasiones surge la duda de si modelar un dato como elemento o como atributo:

```
<precio moneda="euro">100000</precio>
```

o

```
<precio>
    <valor>100000</valor>
    <moneda>euro</moneda>
</precio>
```

- En algunos casos es difícil saber qué opción es preferible
- Se pueden establecer algunas situaciones generales en las que es preferible una alternativa u otra

Elementos vs Atributos

- Utilizar un elemento si:
  - El dato es complejo y puede descomponerse en elementos más simples
  - Puede haber más de un elemento del mismo tipo (ej. Varios teléfonos)
  - El valor del dato puede ser largo
  - El valor del dato cambia con frecuencia
- Utilizar un atributo si:
  - El dato es simple
  - El dato puede tener sólo un número pequeño de valores posibles
- Otro criterio es considerar que los atributos no son contenido
  - ¿Si quitamos todas las marcas (incluidos atributos) el documento sigue teniendo sentido?

Espacios en blanco

- En XML se define como espacio en blanco los siguientes
  - Tabulador \t
  - Avance de línea \n
  - Retorno de carro \r
  - Espacio en blanco \s
- Un analizador XML debe pasar a la aplicación XML todos los espacios en blanco que aparecen dentro del contenido de un documento
- Un analizador XML debe eliminar todos los espacios en blanco de las etiquetas y de los valores de los atributos
- Los analizadores convierten todos los caracteres de fin de línea en caracteres de avance de línea

Referencia a entidades

- Permiten insertar una cadena de caracteres en el contenido de un elemento o en el valor de un atributo
- Hay cinco entidades predefinidas en XML:
  - < (<)
  - & (&)
  - > (>)
  - ' (‘)
  - " (“)
- También es posible crear entidades definidas por el usuario

Otros componentes

- Comentarios
  - Se marcan entre `<!- y -->`
  - Pueden aparecer en cualquier parte del documento fuera de otra marca

```
<!-- un comentario -->
```

- Secciones CDATA
  - Pueden aparecer en cualquier parte del documento donde puedan aparecer datos formados por caracteres
  - No pueden anidarse
  - Se usan para introducir texto que de otra manera sería reconocido como marcado
  - Se marcan entre '<![CDATA[' y ']]>'

```
<![CDATA[<no marcado>]]>
```

  - Ejemplo:

```
<ejemplo>
   <![CDATA[
      <HTML>
       <HEAD><TITLE>Rock & Roll</TITLE></HEAD>
   ]]>
</ejemplo>
```

Sin utilizar CDATA:

```
<ejemplo>
   &lt;HTML>
   &lt;HEAD>&lt;TITLE>Rock &amp; Roll&lt;/TITLE>&lt;/HEAD>
</ejemplo>
```

Documentos bien formados

- Documentos que cumplen las reglas básicas de XML de modo que pueden ser procesados por un programa

- Se dice que un documento XML está **bien formado** cuando cumple las siguientes reglas:
  - El documento debe tener exactamente un elemento de nivel superior (elemento documento o elemento raíz)
    - La primera marca del documento es la marca del elemento raíz
  - Los elementos deben estar adecuadamente anidados, de modo que el contenido tenga una estructura de árbol
  - Cada elemento debe tener una marca de inicio y una marca de fin
    - Los elementos vacíos pueden indicarse con una marca especial
  - El nombre del tipo de elemento de una marca de inicio debe corresponder exactamente con su marca de fin correspondiente
    - En los nombres de los tipos de elementos se distingue entre mayúsculas y minúsculas
  - No puede aparecer un atributo más de una vez, en un mismo elemento
  - El valor de los atributos debe ir entre comillas

Espacios de nombres: XML namespaces

- En ocasiones puede interesar utilizar en un mismo documento elementos y atributos pertenecientes a distintos vocabularios XML
- Esto puede conducir a un problema: la colisión de nombres
  - Podría suceder que elementos (o atributos) correspondientes a vocabularios diferentes tuviesen el mismo nombre
- Para evitar este problema se utilizan los espacios de nombres
  - La idea fundamental es cualificar el nombre de elementos y atributos con un prefijo correspondiente al espacio de nombres de su vocabulario, de forma que el nombre completo resultante sea único

Espacios de nombres: XML namespaces

- Declaración de espacios de nombres
  - Se utiliza un atributo especial: xmlns:prefijo, que especifica el prefijo que precederá a los nombres de ese espacio de nombres
  - Si no se indica ningún prefijo, xmlns declara el espacio de nombres predeterminado
  - El atributo se puede definir en cualquier elemento. Define el uso de un espacio de nombres dentro de ese elemento y sus hijos
  - El valor del atributo de xmlns debe ser una URI válida

```
<elemento xmlns:prefijo="URI">
<elemento xmlns="URI">
```

- La URI que se especifica como valor del atributo xmlns sólo sirve como identificador único del espacio de nombres
  - No referencia a ningún tipo de declaración, ni su contenido se utiliza para ningún tipo de validación

Ejemplos de espacios de nombres

```
<b:inversiones
      xmlns:b=”http://www.bolsa.com”
      xmlns:g=”http://www.geografia.es” >
  <g:pais g:nombre=”Francia”>
      <g:capital> Paris </g:capital>
      <b:capital> 1200 </b:capital>
  </g:pais>
</b:inversiones>
```

Bibliografía

- W3C: http://www.w3.org/TR/1998/REC-xml-19980210
- Curso "XML: Lenguaje de Marcas Extensible“
  Autor: Bartolomé Sintes Marco
- Libro: Lenguajes de Marcas y Sistemas de Gestión de la Información

  Autor: J.M. Castro Ramos

  Editorial: Garceta
- Introducción a XML

  Autor: Alfredo Reino