U.T. II.- Lenguaje de Marcas Extendido: XML.

Índice

1. Introducción.

2. Conceptos y vocabulario.

3.- Elementos.

4.- Instrucciones de Procesamiento.

5.- Comentarios y secciones CDATA.

6.- Referencias de Carácter y de Entidad.

7.- Documentos XML Bien Formados.

8.- Estructura de un documento XML.

9.- Recomendaciones XML

1. Introducción.

- **XML**, acrónimo de e**X**tensible **M**arkup **L**anguage, es un formato estándar diseñado por el W3C a partir de SGML (S*tructured Generalized Markup Language*) para representar datos estructurados de forma jerárquica, es decir, en forma de árbol.
- Aunque, a primera vista, un documento XML puede parecer similar a un HTML, existe una diferencia fundamental: mientras HTML contiene sólo texto, los documentos XML contienen fundamentalmente datos autodefinidos.
- Entre sus ventajas se encuentra su aceptación casi universal, su legibilidad y su carácter autocontenido.
- Su mayor desventaja deriva precisamente de dicho carácter autocontenido, al tener que ir acompañado cada dato de sus correspondientes metadatos, los documentos XML alcanzan un gran tamaño.

- En la actualidad, XML tiene gran número de aplicaciones.
    - La mayoría de los portales y sitios de noticias están basados en XML, dado que permite almacenar la información con una determinada estructura y posteriormente aplicarle fácilmente transformaciones para su presentación.
    - Un proveedor de contenidos, que tendrá la información contenida en una BD, almacenará dicha información en un documento XML para enviarla al cliente, que a su vez transformará el documento recibido al formato para él apropiado.
        - Podrá, por ejemplo, transformarlo en un script SQL para almacenar la información recibida en su propia BD.
        - También podría transformarlo en un documento HTML para mostrar la información que contiene por Internet.

Ejemplo 1:

Los datos referentes a un pedido, que en un documento de texto se almacenarían como:

P-123,C-45,R-567,40,31.22

En un documento XML se almacenarían:

```
<pedido>
        <id>P-123</id>
        <cliente>C-45</cliente>
        <producto>R-567</producto>
        <cantidad>40</cantidad>
        <precio>31.22</precio>
</pedido>
```

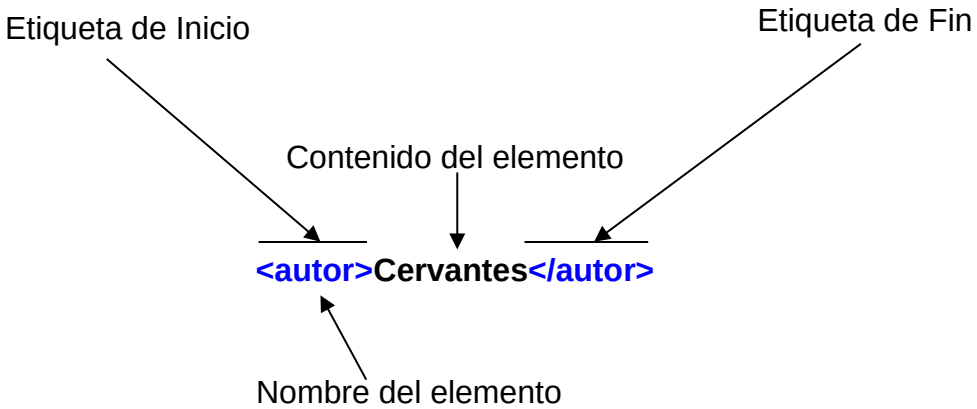
2. Conceptos y vocabulario.

- **Documento XML**: Un documento XML es un documento de texto plano (sin formato) que cumple las reglas de sintaxis de la recomendación XML, se dice entonces que se trata de un documento XML "bien formado".

- **Procesador XML** (XML parser): Cuando una aplicación necesita leer un documento XML, la aplicación recurre a un procesador XML.
    - El procesador o intérprete XML lee el documento, analiza el contenido y le pasa la información en un formato estructurado a la aplicación.
- **Juego de caracteres:** Los documentos XML pueden estar codificados en distintos juegos de caracteres (ISO-8859-1, UTF-8, etc.).
- **Marcas y contenido:** 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.

3.- Elementos.

- En un documento XML el elemento es la unidad básica de información.
- Un elemento se define con sus etiquetas de inicio y fin, ambas obligatorias.
    - Las etiquetas van encerradas entre los caracteres **<** y **>**.



- Un **nombre de elemento** debe comenzar con letra o subrayado bajo, después puede escribirse cualquier conjunto de caracteres excepto el

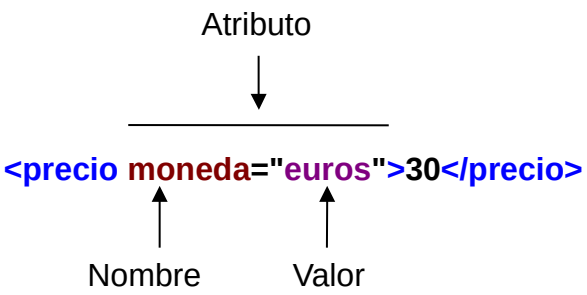
espacio en blanco, los dos puntos, los símbolos de mayor y menor y la barra ***/***.

- Un nombre de elemento no puede comenzar con las letras **xml**.

- Un elemento puede contener a su vez a otros elementos.

3.1.- Atributos.

- Los elementos de un documento XML pueden incluir atributos que los describan.
  - Por ejemplo el tipo de datos que contiene, los valores permitidos, etc.
- Los atributos de un elemento deben especificarse en su etiqueta de inicio.
- Dentro de un mismo elemento no puede repetirse el mismo atributo.
- Un atributo está compuesto por un nombre y un valor.
- El valor deberá ir entre comillas dobles o simples en función del contenido.
  - Si en la información hay comillas dobles el valor se enmarcará con simples y viceversa.



3.2.- Elementos Vacíos.

- Un documento XML puede incluir elementos vacíos, es decir, sin contenido.
- En este caso puede sustituirse la etiqueta de fin de elemento por la barra **/** al final de la etiqueta de inicio.

```
<telefono preferente="sí" />
```

Sería equivalente a:

```
<telefono preferente="sí" ></telefono>
```

4.- Instrucciones de Procesamiento.

- Dentro de un documento XML, generalmente al comienzo, puede incluirse información sobre el resto del documento, i.e., metainformación, mediante lo que se denominan **instrucciones de procesamiento**.
- Una instrucción de procesamiento debe ir encerrada entre los caracteres **<?** y **?>**
- Las instrucciones de procesamiento se utilizan para indicar el sistema de codificación empleado, la hoja **XSLT** que se utilizará para visualizar el documento, etc.
- La única instrucción de procesamiento obligatoria es la que especifica que se trata de un documento XML y su versión.

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

- También deberá especificarse el sistema de codificación empleado si no es Unicode:

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

  - ,Si no se especifica, se asume por defecto el sistema Unicode **"UTF-8"**

5.- Comentarios y secciones CDATA.

- Pueden incluirse **comentarios** en cualquier lugar de un documento XML delimitados por los caracteres **<!--** y **-->**
- También puede incluirse cualquier tipo de contenido dentro de una sección **CDATA**.
    - las secciones CDATA vienen delimitadas por las siguientes secuencias de caracteres: **<![CDATA[** y **]]>**
    - Los analizadores XML consideran el contenido de una sección **CDATA** como comentario y no lo interpretan.

6.- Referencias de Carácter y Entidades.

- Existen algunos caracteres reservados que no podemos utilizar en un documento XML, dado que forman parte de su sintaxis, es decir, tienen un significado propio para el analizador.
- Si queremos incluir alguno de estos caracteres en un documento XML debemos utilizar las denominadas "Referencias de carácter" que no son más que secuencias de caracteres que sustituirán en el documento a los caracteres reservados.
- En **XML** se definen las siguientes referencias de carácter:

Carácter reservado	Referencia de carácter
<	&lt;
>	&gt;
&	&amp;
'	&apos;
"	&quot;

- XML también soporta referencias de carácter con la sintaxis:

  &#nnn;

  - Donde **nnn** es el número decimal Unicode del carácter a representar.
  - Por ejemplo, el símbolo @ se escribiría **@**.

- Una "**entidad**" consiste en un nombre y su valor.
  - Las entidades se definen en el prólogo de un documento XML mediante la etiqueta <!ENTITY nombre "valor">
  - Una vez definida una entidad, haremos referencia a ella en el cuerpo del documento escribiendo su nombre precedido del carácter **&** y seguido del carácter **;**.
  - Ejemplo:

```
<!ENTITY IGN "Instituto Geográfico Nacional">
.....
<organismo>&IGN;</organismo>
```

  - Sería el equivalente a la definición de una constante en un lenguaje de programación.

7.- Documentos XML Bien Formados.

- Un documento XML consta de un prólogo y un elemento raíz.
- El prólogo contiene información sobre el resto del documento.
  - En él se escriben las instrucciones de procesamiento, secciones CDATA, entidades, así como una descripción de la estructura del documento, generalmente recogida en una DTD (Document Type Definition) o en un Schema XML.
- 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 debe cumplir los siguientes puntos:

1. El documento contiene únicamente caracteres válidos.
2. Hay un elemento raíz que contiene al resto de elementos.
3. Los nombres de los elementos y de sus atributos no contienen espacios.
4. El primer carácter de un nombre de elemento o de atributo debe ser una letra, dos puntos (**:**) o subrayado (_).
5. El resto de caracteres pueden ser también números, guiones (**-**) o puntos (**.**).
6. Los caracteres "**<**" y "**&**" sólo se pueden utilizar como comienzo de marcas.
7. 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.
8. Las etiquetas de cierre coinciden con las de apertura (incluso en el uso de mayúsculas y minúsculas).
9. Las etiquetas de cierre no contienen atributos.
10. Ninguna etiqueta tiene dos atributos con el mismo nombre.
11. Todos los atributos tienen algún valor.
12. Los valores de los atributos están entre comillas.
13. No existen referencias en los valores de los atributos.

Ejercicio 201.

Crear con **XML Copy Editor** los siguientes documentos XML, comprobar si están bien formados y abrirlos con el navegador.

a)

```
<?xml version="1.0" encoding="utf-8"?>
<libro>
        <autor>Isabel Castro</autor>
        <titulo>XML Guía de Aprendizaje</titulo>
        <precio moneda="euros">30</precio>
</libro>
```

b)

```
<?xml versión="1.0" encoding="ISO-8859-1"?>
 <agenda>
   <entrada>
         <nombre-completo>Marta Elena Tablada</nombre-completo>
         <direccion>
            <calle> Avda Los Castros</calle>
            <ciudad>Santander</ciudad>
            <codigo-postal>39005</codigo-postal>
            <region>Cantabria</region>
            <pais>España</pais>
         </direccion>
         <tel preferente="true"> 942201363</tel>
         <correo-e href="tabladam@unican.es"/>
      </entrada>
   </agenda>
```

- Como explicación del punto 10 de la recomendación XML, si el documento del ejercicio 1 a) lo hubiéramos escrito:

```
<?xml version="1.0" encoding="utf-8"?>
<libro
      autor="Isabel Castro"
      titulo="XML Guía de Aprendizaje"
      precio="30€"
</libro>
```

Estaría bien formado, pero no podríamos incluir libros con más de un autor.

Ejercicio 202.

Modificar el ejercicio 201 a) para añadir como autor a "Domingo Fernández Pérez".

Ejercicio 203.
Los siguientes documentos están mal formados. ¿Por qué?

a)
```
<?xml version="1.0" encoding="ISO-8859-1"?>
<agenda>
<entrada>
<nombre-completo>Marta Elena Tablada
</entrada>
</agenda>
```

b)
```
<?xml version="1.0" encoding="ISO-8859-1"?>
<nombre-completo>Marta Elena Tablada
</nombre-completo>
<direccion>
<calle> Avda Los Castros</calle>
<ciudad>Santander</ciudad>
<codigo-postal>39005</codigo-postal>
<region>Cantabria</region>
<pais>España</pais>
</direccion>
<tel preferente="true"> 942201363</tel>
<correo-e href="tabladam@unican.es"></correo-e>
</direccion>
```

c)
```
<?xml version="1.0"?>
<padre>
<hijo1>
<nombre>Juan<apellido>Pérez</apellido></nombre>
</hijo1>
<hijo2>
<nombre>Luis<apellido>García</nombre></apellido>
</hijo2>
</padre>
```

d)
```
<?xml version="1.0"?>
<hijo1>
<nombre>Juan</nombre>
<apellido>Pérez</apellido>
</hijo1>
<hijo2>
<nombre>Luis</nombre>
<apellido>García</apellido>
</hijo2>
```

e)

```
<?xml version="1.0" encoding="UTF-8"?>
<pelicula>
<titulo>Con faldas y a lo loco</titulo>
<director>Billy Wilder</director>
</pelicula>
<pelicula>
<director>Leo McCarey</director>
<titulo>Sopa de ganso</titulo>
</pelicula>
<autor />barto</autor>
```

f)

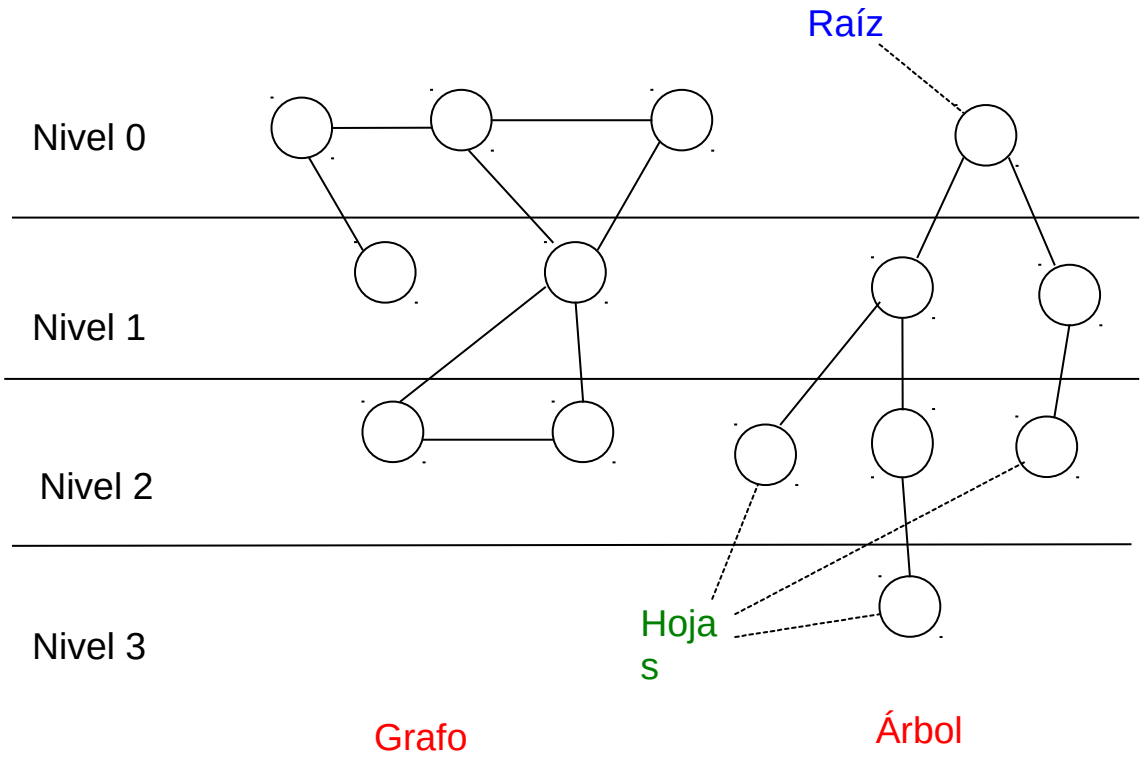
```
<?xml version="1.0" encoding="UTF-8"?>
<deportistas>
<deportista>
<deporte Atletismo />
<nombre>Jesse Owens</nombre>
<deportista>
<deporte Natación />
<nombre>Mark Spitz</nombre>
</deportista>
</deportistas>
```

8.- Estructura de un documento XML.

- Como ya hemos dicho, todos los elementos de un documento XML bien formado deben seguir una estructura estrictamente jerárquica (en forma de árbol).
- Desde un punto de vista matemático, un árbol es un caso particular de grafo.
- Un grafo es un conjunto de objetos, llamados nodos, agrupados en niveles y relacionados entre sí.
    - Dado un nodo cualquiera, llamamos nodos **padre** a los nodos de nivel inmediatamente superior relacionados con él.
    - Llamamos nodos **hijo** a aquellos nodos de nivel inmediatamente inferior relacionados con él.
    - Sus nodos **hermano** serían aquellos nodos de su mismo nivel relacionados con él.

- Llamamos nodos **raíz** a los nodos de nivel superior.
- Llamamos nodos **hoja** a los nodos que no tienen hijos.

- Un árbol es un grafo que cumple las siguientes condiciones:
  - Tiene un único nodo raíz.
  - Cada nodo, excepto el raíz, tiene un único padre.
  - Cualquier nodo puede tener 0 ó más hijos.
  - Ningún nodo puede tener hermanos.

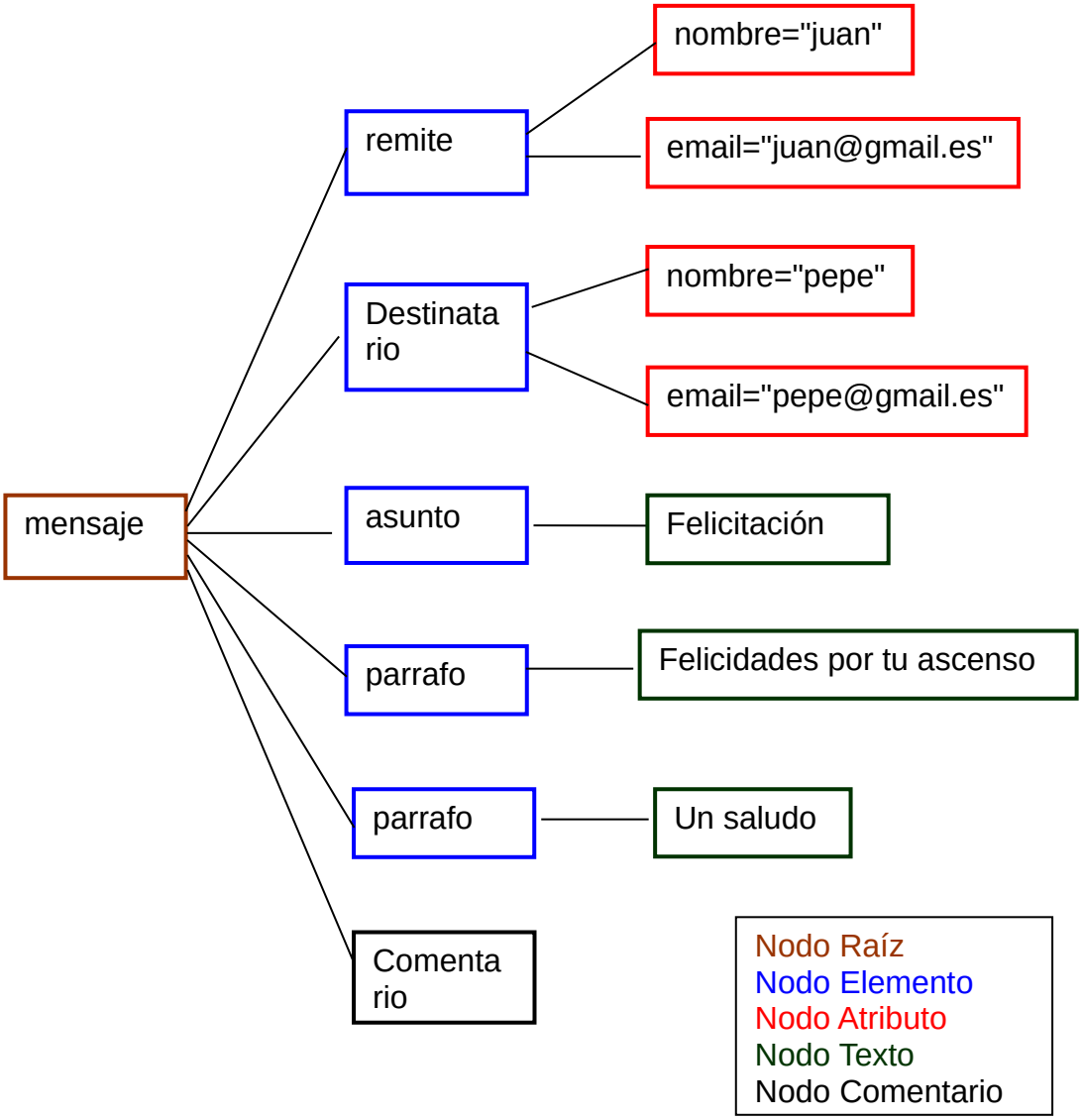


- De esta manera, los analizadores de documentos XML tratan a éstos como estructuras en árbol, considerando que cada elemento es un nodo, que un elemento es padre de los elementos que contiene e hijo del elemento en el que está contenido.

8.1- Tipos de nodos.

- Aunque la especificación completa XML define 12 tipos de nodos, los más habituales son:
    - **Nodo Raíz**. De él derivan todos los demás nodos del árbol.
    - **Nodo Elemento**. Único nodo que puede contener atributos y del que se pueden derivar otros nodos.
    - **Nodo Atributo**. Contiene un par nombre/valor.
    - **Nodo Texto**. Contiene el texto contenido en un elemento.
    - **Nodo Comentario**. Contiene un comentario.
    - También se especifican tipos de nodos para instrucciones de procesamiento, espacios de nombres, etc.

Ejercicio 204: Dado el siguiente árbol de nodos:



Escribir el documento XML que representa.

Ejercicio 205.
Dado el siguiente documento XML, crear un árbol para representar su estructura.

```
<?xml version="1.0"?>
<libro>
<titulo>El Ingenioso Hidalgo Don Quijote de la Mancha</titulo>
 <autor>Miguel de Cervantes Saavedra</autor>
<fecha>
<publicacion>1605</publicacion>
<edicion>2009</edicion>
</fecha>
```

<localizacion>
<estanteria>B</estanteria>
<fila>7</fila>
</localizacion>
</libro>

Ejercicio 206.

Dado el siguiente documento:

FACTURA Núm. 999			
CLIENTE Número: 879 Nombre: José Rodríguez Teléfono: 910101010			
Datos Factura			
Ref.	Descripción	Cantidad	Precio
A101	Placa Base QDI	1	230
A105	DIMM DDR 512Mb	4	110
A107	Disco Duro 800Gb	1	150

Se pide:
1. Estructura en árbol con su contenido.
2. Documento XML correspondiente.
Nota: los campos Núm, Número y Ref deben representarse como atributos.

9.- 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: