Ud 1. Introducción a los lenguajes de marcas.

RA1. Interpreta lenguajes de marcas reconociendo sus principales características e identificando sus elementos.

Criterios de evaluación:
a) Se han identificado las características generales de los lenguajes de marcas.
b) Se han reconocido las ventajas que proporcionan en el tratamiento de la información.
c) Se han clasificado los lenguajes de marcas e identificado los más relevantes.
d) Se han diferenciado sus ámbitos de aplicación.
e) Se ha reconocido la necesidad y los ámbitos específicos de aplicación de un lenguaje de marcas de propósito general.
f) Se han analizado las características propias del lenguaje XML.
g) Se ha identificado la estructura de un documento XML y sus reglas sintácticas.
h) Se ha contrastado la necesidad de crear documentos XML bien formados y la influencia en su procesamiento.
i) Se han identificado las ventajas que aportan los espacios de nombres.

Contenido:

Introducción
Definición
Características.
Clasificación.
Ámbito de uso.
XML. Estructura y sintaxis.

Introducción.

En los años 70 continúa la evolución de los **lenguajes de programación** a la vez que surgen otros lenguajes informáticos orientados a la **gestión de información**. Con el desarrollo de los programas que procesan texto (editores y procesadores de texto que agilizan el proceso de edición) surgen los primeros lenguajes informáticos especializados en tareas de descripción y estructuración de información: los lenguajes de marcas.

Los lenguajes de marcas surgieron, inicialmente, como lenguajes formados por el conjunto de códigos de formato que los procesadores de texto introducen en los documentos para dirigir el **proceso de presentación** (impresión) mediante una impresora. Como en el caso de los lenguajes de programación, inicialmente estos códigos de formato estaban ligados a las características de una máquina, programa o procesador de textos concreto y, en ellos, inicialmente no había nada que permitiese al programador (formateador de documentos en este caso) abstraerse de las características del procesador de textos y expresar de forma independiente a éste la estructura y la lógica interna del documento.

Este marcado estaba exclusivamente orientado a la presentación de la información, aunque pronto fueron conscientes de las posibilidades del marcado y se le dieron nuevos usos que resolverían una gran variedad de necesidades, apareció el formato generalizado.

Uno de los problemas presentes en el desarrollo de este tipo de lenguajes era la **falta de estandarización**. Para resolverlo, en los años sesenta IBM encargó a Charles F. Goldfarb, la construcción de un sistema de edición, almacenamiento y búsqueda de documentos legales. Tras analizar el funcionamiento de la empresa llegaron a la conclusión de que para realizar un buen procesado informático de los documentos había que establecer un **formato estándar** para todos los documentos que se manejaban en la empresa. Con ello se lograba gestionar cualquier documento en cualquier departamento y con cualquier aplicación, sin tener en cuenta dónde ni con qué se generó el documento. Dicho formato tenía que ser válido para los distintos tipos de documentos legales que utilizaba la empresa, por tanto, debía ser flexible para que se pudiera ajustar a las distintas situaciones.

1Charles F. Goldfarb

El formato de documentos que se creó como resultado de este trabajo fue **GML**, cuyo objetivo era describir los documentos de tal modo que el resultado fuese independiente de la plataforma y la aplicación utilizada.

El formato GML evolucionó hasta que en 1986 dio lugar al **estándar ISO 8879** que se denominó **SGML**.

SGML introduce tres conceptos básicos:

- El concepto de lenguaje de marcado generalizado como un metalenguaje que sirve para **definir lenguajes** concretos que pueden adaptarse a cada dominio mediante una gramática que describe formalmente un tipo específico de documento o **DTD** (Document Type Definition);
- El concepto de **marcado descriptivo**, frente marcado procedural. El marcado descriptivo describe, mediante las marcas o etiquetas definidas en la DTD, la estructura lógica de la información. La idea clave es que las marcas no determinan el procesamiento del documento de manera fija, ya que dicho procesamiento se determina a partir de las necesidades concretas, y se beneficia de la estructura lógica del documento caracterizada a través de sus marcas;
- El concepto de **independencia de la plataforma**. Como los documentos SGML únicamente contienen texto, éstos pueden ser procesados en distintas plataformas, trascendiendo el uso de dichos documentos a los sistemas que los crearon y utilizaron originariamente.

En 1989/90 Tim Berners-Lee se encontró con la necesidad de organizar, enlazar y compatibilizar gran cantidad de información procedente de diversos sistemas. Conociendo SGML creó a partir de su sintaxis un lenguaje de descripción de documentos llamado HTML como combinación de dos estándares.

- ASCII: Es el formato que cualquier procesador de textos sencillo puede reconocer y almacenar. Por tanto, es un formato que permite la trasferencia de datos entre diferentes ordenadores.
- SGML: Lenguaje que permite dar estructura al texto, resaltando los títulos o aplicando diversos formatos al texto.

HTML es una versión simplificada de SGML, ya que sólo se utilizaban las instrucciones absolutamente imprescindibles. Era tan fácil de comprender que rápidamente tuvo gran aceptación. Se crea la World Wide Web y HTML se convierte en el lenguaje para la creación de sus documentos.

En 1998 surge el estándar internacional **XML**, un lenguaje de marcas puramente estructural que no incluye ninguna información relativa al diseño, que permite crear etiquetas adaptadas a las necesidades (de ahí lo de "extensible"). Está convirtiéndose con rapidez en estándar para el **intercambio de datos** en la Web. A diferencia de HTML las etiquetas indican el significado de los datos en lugar del formato con el que se van a visualizar los datos.

Definición.

Un lenguaje de marcado o lenguaje de marcas es un LENGUAJE que incorpora en el contenido un conjunto de etiquetas o marcas que añaden al documento información adicional acerca de su estructura, presentación, etc... El lenguaje de marcas es el que especifica cuáles serán las etiquetas posibles, donde deben colocarse y el significado que tendrá cada una de ellas. Permiten hacer explícita la estructura de un documento, su contenido semántico o cualquier otra información lingüística o extralingüística que se quiera hacer patente.

No hay que confundir lenguaje de marcas con lenguaje de programación.

- ¿Conoces algún lenguaje de programación.

Ejemplos:

```
<p>
   Esto es un texto escrito en HTML
</p>
```

```
<noticia>
   <fecha>20/09/2020</fecha>
   <lugar>Córdoba</lugar>
   <titular>Empieza el curso</titular>
   <desarrollo>Se presenta un . . .</desarrollo>
</noticia>
```

Características.

- Se crean con editores de texto, usando archivos de texto plano.
- Permiten la utilización de metadatos.
- Utilizan marcas para incorporar los metadatos al documento.
- Compacidad. Las marcas aparecen junto al contenido del documento.
- Fáciles de interpretar y procesar.
- Flexibilidad.

Clasificación.

Aunque en la práctica, en un mismo documento pueden combinarse varios tipos diferentes de lenguajes de marcas, éstos se pueden clasificar en tres grupos:

Orientados a la presentación: Son los utilizados generalmente por los procesadores de texto y codifican cómo ha de presentarse el documento, es decir, define el formato del texto. Por ejemplo: indicando que una determinada palabra se debe presentar en negrita, o que se debe dejar un espacio entre caracteres determinado. Generalmente, las marcas se ocultan al usuario, lo que permite obtener un efecto WYSIWYG . RTF (Rich Text Format) es un tipo de lenguaje de marcado de este tipo.

De procedimiento: Orientados también a la presentación, pero en este caso, dentro de un marco procedural que permite definir macros (secuencias de acciones), es decir, el programa que representa el documento debe interpretar el código en el mismo orden en que aparece. Entre los ejemplos más comunes, encontramos TeX, LaTeX y Postcript.

Descriptivos o semánticos: Este tipo no define qué se debe hacer con un trozo o sección del documento sino que, por el contrario, las marcas sirven para indicar qué es esa información, es decir, describen qué es lo que se está representando. Describen las diferentes partes, pero sin especificar cómo deben representarse. XML, SGML y sus derivados son ejemplos de lenguajes descriptivos.

Ámbitos de uso.

Algunos ejemplos de lenguajes de marcado agrupados por su ámbito de utilización son:

Documentación electrónica

- RTF (Rich Text Format): Formato de Texto Enriquecido, fue desarrollado por Microsoft en 1987. Permite el intercambio de documentos de texto ente distintos procesadores de texto.
- TeX: Su objetivo es la creación de ecuaciones matemáticas complejas.
- Wikitexto: Permite la creación de páginas wiki en servidores preparados para soportar este lenguaje.
- DocBook: Permite generar documentos separando la estructura lógica del documento de su formato. De este modo, dichos documentos, pueden publicarse en diferentes formatos sin necesidad de realizar modificaciones en el documento original.

Tecnologías de internet

- HTML, XHTML: (Hypertext Markup Language, eXtensible Hypertext Markup Language): Su objetivo es la creación de páginas web.
- RSS: Permite la difusión de contenidos web.
- SVG: Gráficos vectoriales.

Otros lenguajes especializados

- MathML (Mathematical Markup Language): Su objetivo es expresar el formalismo matemático de tal modo que pueda ser entendido por distintos sistemas y aplicaciones.
- VoiceXML (Voice Extended Markup Language) tiene como objetivo el intercambio de información entre un usuario y una aplicación con capacidad de reconocimiento de habla.
- MusicXML: Permite el intercambio de partituras entre distintos editores de partituras.

¿Puedes dar más ejemplos de lenguajes de marcas?

Ejemplos:

roff

```
.TL
The Road
.AU
Yours Truly
.AI
Koi Thikana Kahin Nahi
.sp 0.3i
.LP
.QS
ROAD, n. A strip of land along which one may pass from where
it is too tiresome to be to where it is futile to go.
.rj
\[em] Ambrose Bierce, \c
```

SGML

```
<!doctype linuxdoc system>

<article>

<titlepag>
  <TITLE>Ejemplo de formato SGML</TITLE>
  <author>&copy; 2010 Alberto Molina Coballes,
  <tt/alberto@gonzalonazareno.org/
    <abstract>
    Resumen del documento ...
    </abstract>
</titlepag>
<toc>
```

LaTeX

```
\documentclass{article}

\title{Ejemplo de formato LaTeX}
\author{\copyright 2010 Alberto Molina Coballes}
\email{alberto@gonzalonazareno.org}
\begin{abstract}
  Resumen del documento ...
\end{abstract}

\tableofcontents
```

XML. Estructura y sintaxis.

XML, siglas de extensible markup language, es un metalenguaje de marcas desarrollado por el World Wide Web Consortium (**W3C**) , que es un consorcio internacional que crea recomendaciones para la World Wide Web.

XML proporciona una serie de reglas para que cualquiera pueda definir un lenguaje, describiendo su propio conjunto de etiquetas, atributos y relaciones entre estas etiquetas. Al conjunto de reglas de un lenguaje se le llama **gramática del lenguaje**.

Así, XML es un lenguaje de marcas puramente estructural que no incluye ninguna información relativa al diseño. A diferencia de HTML las etiquetas indican el significado de los datos en lugar del formato con el que se van a visualizar los datos, XML es un lenguaje simple de descripción de información.

Características.

- Permitir definir etiquetas propias.
- Permitir asignar atributos a las etiquetas.
- Utilizar un esquema para definir de forma exacta las etiquetas y los atributos.
- La estructura y el diseño son independientes.

Estructura.

Todo documento XML tiene una estructura jerárquica arborescente y está compuesta por dos partes fundamentales: prólogo, y cuerpo.
El **prólogo** es la primera parte del documento y contiene la información (meta información) sobre el resto del documento, como versión XML y juego de caracteres utilizado. También puede incluir la referencia a la gramática, recogida en un DTD (Document Type Definition) o en un Schema XML.
El **cuerpo** del documento se compone de un conjunto de elementos donde cada elemento está formado por:

Etiqueta de inicio. Delimitada por los caracteres < >, incluye nombre de elemento y conjunto de pares atributo, valor.
Contenido. Puede ser texto u otros elementos.
Etiqueta final. Nombre del elemento delimitado por los caracteres </ >

```
<?xml version=”1.0” encoding=”iso-8859-1”
<cine>
   <pelicula estreno=”1999”>
       <titulo> The Matrix</titulo>
       <direccion> Hermanas Wachowski</direccion>
       <categoria> Ciencia ficcion</categoria>
   </pelicula>
</cine>
```

Para que un documento XML se considere correcto debe estar **bien formado** y ser **válido**.

Documentos bien formados: Un documento XML se considera bien formado si cumple las siguientes características o reglas sintácticas:

- Aunque no es obligatorio, W3C recomienda empezar con una declaración XML en la que se indique la versión y el juego de caracteres empleado.
- Estructura jerárquica arborescente. Tiene un único elemento raíz.
- Los elementos deben anidarse correctamente.
- Las etiquetas de cierre de los elementos son obligatorias.
- Se distinguen las mayúsculas y las minúsculas.
- Se permiten elementos vacíos.
- Los valores de los atributos van entrecomillados.
- Los nombres de los atributos deben ser identificadores válidos.
- Los comentarios se deben escribir entre <!-- y -->.

Documentos válidos: Un documento XML se considera válido si está bien formado y además cumple la gramática definida para el lenguaje. La gramática del lenguaje se puede describir por medio de DTD, XML Schema o más recientemente Relax NG.

Espacios de nombres en XML.

Un espacio de nombres XML es una recomendación W3C para proporcionar elementos y atributos con nombre único en un archivo XML. Un archivo XML puede contener nombres de elementos o atributos procedentes de más de un vocabulario XML. Si a cada uno de estos vocabularios se le da un espacio de nombres, un ámbito semántico propio, referenciado a una URI donde se listen los términos que incluye, se resuelve la ambigüedad existente entre elementos o atributos que se llamen igual, la homonimia. Los nombres de elementos dentro de cada espacio de nombres deben ser únicos.

Un ejemplo sería una instancia XML que contuviera referencias a un cliente y a un producto solicitado por este. Tanto el elemento que representa el cliente como el que representa el producto pueden tener un elemento hijo llamado "numero_ID". Las referencias al elemento "numero_ID" podrían ser ambiguas, salvo que los elementos, con igual nombre pero significado distinto, se llevaran a espacios de nombres distintos que los diferenciaran.

```
<?xml version="1.0"?>
<cli:cliente xmlns:cli='http://es.wikipedia.org/wiki/Espacio_de_nombres_XML/cliente'
             xmlns:ped='http://es.wikipedia.org/wiki/Espacio_de_nombres_XML/pedido'>
    <cli:numero_ID>1232654</cli:numero_ID>
    <cli:nombre>Fulanito de Tal</cli:nombre>
    <cli:telefono>99999999</cli:telefono>
    <ped:pedido>
      <ped:numero_ID>6523213</ped:numero_ID>
      <ped:articulo>Caja de herramientas</ped:articulo>
      <ped:precio>187,90</ped:precio>
    </ped:pedido>
</cli:cliente>
```

El espacio de nombres se declara usando el atributo XML reservado xmlns, cuyo valor debe ser un identificador uniforme de recurso.

```
xmlns="http://www.w3.org/1999/xhtml"
```

La URI no se lee realmente como una dirección; se trata como una cadena para evitar las ambigüedades en los identificadores del documento.