

Lenguaje de Marcas y Sistemas de Gestión de Información.



UT 04.01- El Lenguaje XML. Conceptos Básicos.

UT 04.01- El Lenguaje XML. Conceptos Básicos.

1.- Introducción a XML.

1.1.- Definición de XML.

- XML es un estándar internacional desarrollado por el **Grupo de Trabajo de XML** (conocido como el Comité de Revisión Editorial de SGML) formado bajo el auspicio del **World Wide Web Consortium (W3C)** en 1996.
- La Recomendación XML dice textualmente:
"El Lenguaje Extensible de Marcas, abreviado XML, describe una clase de objetos de datos llamados documentos XML y parcialmente describe el comportamiento de programas de computador que pueden procesarlos."
- XML permite que el usuario diseñe sus propias marcas dándole el significado que necesita.
- Un fichero XML correctamente escrito tiene dos cualidades necesarias:
 - **Bien formado (well formed).** Se ha escrito de acuerdo con el estándar.
 - **Válido.** Cumpliendo con la definición del estándar está, lógicamente bien estructurado y define en su totalidad cada uno de sus contenidos sin ambigüedad alguna.



UT 04.01- El Lenguaje XML. Conceptos Básicos.

1.- Introducción a XML.

1.2.- Características de XML.



Según la W3C, podemos explicar las características de XML mediante **siete puntos importantes**:

1. Es un estándar para almacenar datos estructurados **en un fichero de texto**.
2. En XML se usan marcas y atributos, para delimitar fragmentos de datos, dejando la interpretación de éstos a la aplicación que los lee.
3. Es **legible de forma informática, no por seres humanos**.
4. XML **es una familia de tecnologías asociadas** al formato.
5. Los ficheros resultantes, son mas grandes que sus equivalentes binarios.
6. Es una **tecnología madura** y en continuo crecimiento.
7. XML **no requiere licencias, es independiente de la plataforma, y tiene un amplio soporte**.

UT 04.01- El Lenguaje XML. Conceptos Básicos.

1.- Introducción a XML.

1.3.- Ejemplo de fichero XML.

- En este ejemplo encontramos los siguientes elementos de marcado, cada uno con su marca inicial y final. Las marcas no están predefinidas, sino que podemos definir nuestras propias marcas que cumplan el propósito buscado.
- XML busca la separación entre el contenido (información del documento), la estructura (tipo y organización de los elementos que componen el documento) y la presentación (manera en que la información es presentada al usuario).

```
<alumnos>
  <alumno id="321">
    <nombre>Juán Martínez</nombre>
    <calificaciones>
      <calificacion modulo="LMSGI">6.50</calificacion>
      <calificacion modulo="PRG">7.25</calificacion>
      <calificacion modulo="FOL">6.75</calificacion>
    </calificaciones>
  </alumno>
  <alumno id="627">
    <nombre>Ana López</nombre>
    <calificaciones>
      <calificacion modulo="LMSGI">7.25</calificacion>
      <calificacion modulo="PRG">6.75</calificacion>
      <calificacion modulo="FOL">5.25</calificacion>
    </calificaciones>
  </alumno>
</alumnos>
```

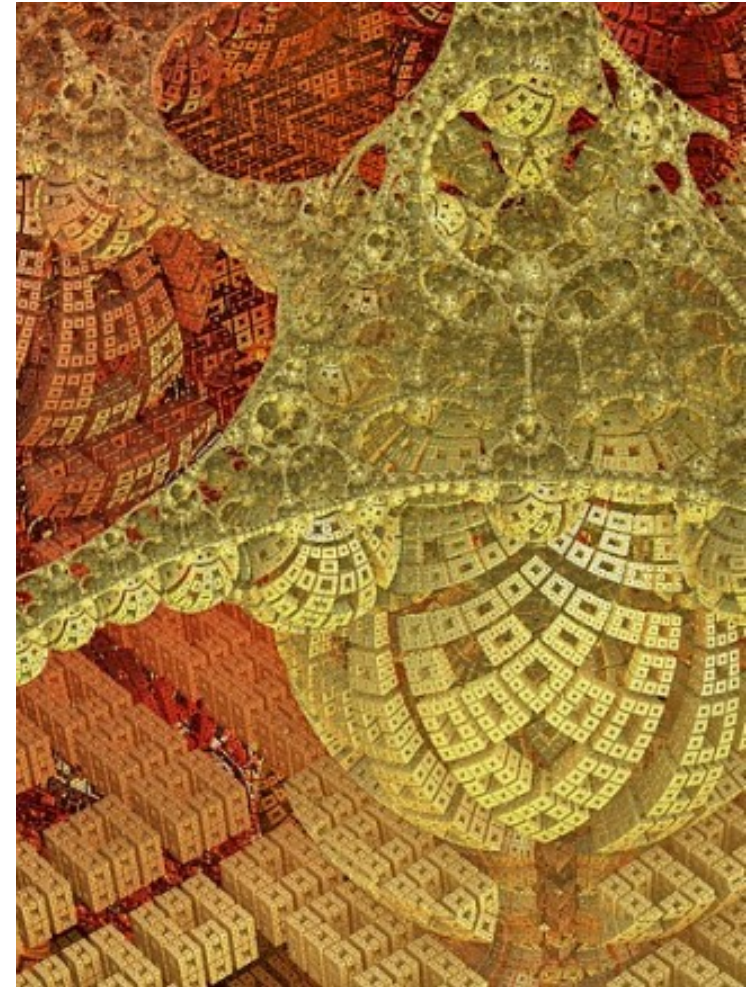
UT 04.01- El Lenguaje XML. Conceptos Básicos.

2.- Estructura de un documento XML.

2.1.- Restricciones sintácticas.

Un documento XML bien formado debe cumplir los siguientes requisitos en su estructura.

1. No se permite la anidación incorrecta de elementos.
2. No se permiten elementos sin etiqueta de cierre. Los elementos que no poseen contenido deben de utilizar una etiqueta de cierre, o usar la abreviatura permitida en XML, consistente en incluir una barra vertical (/) antes del carácter de cierre (>).
3. Todos los atributos deben de ir entre comillas dobles.
4. XML diferencia entre mayúsculas y minúsculas.



UT 04.01- El Lenguaje XML. Conceptos Básicos.

2.- Estructura de un documento XML.

2.2.- Estructura general de un documento XML.

- La estructura general de un documento XML está formada por dos partes:
 - **Prólogo.** Contiene una secuencia de instrucciones de procesamiento y de declaración del tipo de documento. Es de carácter opcional.
 - **Cuerpo.** Contenido de información del documento, organizado como un árbol único de elementos marcados.
- El estándar también permite la inclusión opcional de un **epílogo**, al final del documento, que puede contener instrucciones de procesamiento.
- En el documento siguiente se puede distinguir el **prólogo (rojo)** y el **cuerpo (azul)**.

```
<?xml version="1.0" encoding="UTF-8"?>
<ciudades>
  <ciudad>
    <nombre>Nueva Delhi</nombre>
    <pais continente="Ásia">India</pais>
  </ciudad>
  <ciudad>
    <nombre>Lisboa</nombre>
    <pais continente="Europa">Portugal</pais>
  </ciudad>
  <ciudad>
    <nombre>El Cairo</nombre>
    <pais continente="África">Egipto</pais>
  </ciudad>
</ciudades>
```

UT 04.01- El Lenguaje XML. Conceptos Básicos.

2.- Estructura de un documento XML.

2.3.- Prólogo de un documento XML. Declaración del documento.

- El prólogo identifica al fichero como documento XML e incluye información sobre la versión de XML utilizada para su definición.
- Además, puede incluir información adicional. En esta tabla se muestra ejemplo de parámetros definibles:

Parámetro	Utilidad
version	Permite indicar la versión para la que se elaboró el documento.
encoding	Indica el juego de caracteres utilizado en el documento. Por defecto es UTF-8 , aunque se suele utilizar la codificación de 8 bits ISO-8859-1 , asociada a los lenguajes de Europa Occidental.
standalone	Puede valer tener como valores “yes” o “no”. El valor “yes”, indica que el documento contiene en su interior toda la información relevante para su interpretación.

- Un ejemplo de declaración XML completa podría ser:

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

Declaración del documento.

- Accesoriamente se puede incluir la **declaración del documento**. esta provee una serie de mecanismos que aportan funcionalidad a XML. Definiendo una serie de restricciones adicionales que deben cumplir el documento. Un ejemplo de una declaración de tipo de documento es el siguiente:

`<!DOCTYPE Casas_Rurales SYSTEM “http://www.casasrurales.com/casasrurales.dtd”>`

UT 04.01- El Lenguaje XML. Conceptos Básicos.

2.- Estructura de un documento XML.

2.4.- Cuerpo de un documento XML. Definición.

- El cuerpo es la parte que contiene la información del documento.
- El cuerpo de los documentos XML tiene una estructura de árbol, en la que siempre existe un elemento principal, o **elemento raíz (root)**, dentro del cual se encuentran el resto de los elementos. Todos elementos de un documento XML pueden a su vez contener subelementos o elementos hijos.
- En este ejemplo, <libreria> es la raíz del documento de la que cuelgan <libro> y así sucesivamente.

```
<?xml version="1.0" encoding="UTF-8"?>
<libreria>
  <libro categoría="COOKING">
    <título lang="en">Everyday Italian</título>
    <autor>Giada De Laurentiis</autor>
    <año>2005</año>
    <precio>30.00</precio>
  </libro>
  <libro categoría="INFANTIL">
    <título lang="en">Harry Potter</título>
    <autor>J K. Rowling</autor>
    <año>2005</año>
    <precio>29.99</precio>
  </libro>
  <libro categoría="WEB">
    <título lang="en">Learning XML</título>
    <autor>Erik T. Ray</autor>
    <año>2003</año>
    <precio>39.95</precio>
  </libro>
</libreria>
```


UT 04.01- El Lenguaje XML. Conceptos Básicos.

2.- Estructura de un documento XML.

2.4.- Cuerpo de un documento XML. Juego de caracteres.



- Aunque todos los caracteres de un documento XML deben estar recogidos en los estándares Unicode - ISO/IEC 10646, **no todos los caracteres de estos estándares se permiten en un documento XML.**
- El uso de la mayoría de los caracteres no gráficos está prohibido. No se pueden incluir en un documento XML, ya sea directamente o mediante referencias a carácter.
- Podemos incluir en nuestro documento dichos caracteres mediante **una referencia a carácter**. Esta se compone de la forma siguiente:

- **&#NNNNN;** decimal (hasta 5 dígitos).
- **&#xNNNN;** hexadecimal (hasta 4 dígitos)

El código numérico (decimal o hexadecimal) corresponde al código Unicode.

- Hay caracteres, como “<” y “&”, que no se pueden usar directamente. Por tanto, se introducen como referencias a entidades. Hay 5 entidades predefinidas: **<** (<), **>** (>), **&** (&), **'** (') y **"** (“). Por ejemplo, si se desea en una cadena de texto incluir el símbolo de mayor se introduciría lo siguiente:

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

```
<ejemplo>
```

Esta cadena incluye el símbolo **<**

```
</ejemplo>
```

UT 04.01- El Lenguaje XML. Conceptos Básicos.

3.- Contenido del cuerpo de un documento XML.

3.1.- Elementos de información.

- Son las distintas hojas del árbol de información que contiene el documento.
- En un documento XML siempre habrá como mínimo el elemento raíz.
- Los elementos están formados por tres componentes: etiqueta de inicio o apertura, etiqueta de finalización o cierre y contenido, situado entre ambas etiquetas.

<etiqueta>Contenido</etiqueta>

El hecho de no respetar este requisito implica un error a la hora de validar el documento.

- En XML pueden existir elementos sin contenido, que reciben el nombre de **elementos vacíos**. Se pueden expresar de dos formas:

<etiqueta></etiqueta>

<etiqueta/>



UT 04.01- El Lenguaje XML. Conceptos Básicos.

3.- Contenido del cuerpo de un documento XML.

3.2.- Atributos.



- Los elementos XML pueden incluir, dentro de la etiqueta de inicio, atributos opcionales.
- Los atributos actúan como modificadores que incluyen información adicional aplicable a un elemento.
- **Ejemplo.** Dados los siguientes datos de un producto:
 - Código: G45.
 - Nombre: Gorro de lana.
 - Color: negro.
 - Precio: 12.56

Una posible representación en un documento XML podría ser la siguiente:

```
<?xml version="1.0" encoding="UTF-8"?>
<producto codigo="G45">
  <nombre color="negro" precio="12.56">
    Gorro de lana
  </nombre>
</producto>
```

- Los nombres de los atributos deben cumplir las mismas normas de sintaxis que los nombres de los elementos. Además, todos los atributos de un elemento tienen que ser únicos.

UT 04.01- El Lenguaje XML. Conceptos Básicos.

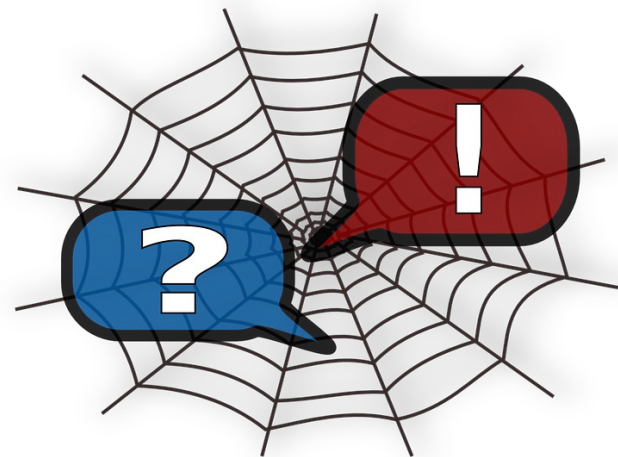
3.- Contenido del cuerpo de un documento XML.

3.3.- Comentarios.

- Los comentarios son una forma de añadir información sobre el documento, que en principio no va a ser utilizada por las aplicaciones informáticas, sino por los programadores.
- Los comentarios no se consideran datos carácter, y serán ignorados por los procesadores XML. Se pueden incluir comentarios en cualquier parte del documento.
- Los comentarios en XML se representan igual que en HTML, es decir:

<!-- ... texto del comentario ... -->

El texto del comentario admite cualquier carácter salvo la cadena "--".



```
<?xml version="1.0" encoding="UTF-8"?>
<!--Ejemplo uso de comentarios.-->
<a>
  <b>
    <c cantidad="4">cccc</c>
    <d cantidad="2">dd</d>
  </b>
  <!--e puede aparecer varias veces.-->
  <e cantidad="5">eeee</g>
  <e cantidad="2">ee</g>
</a>
```

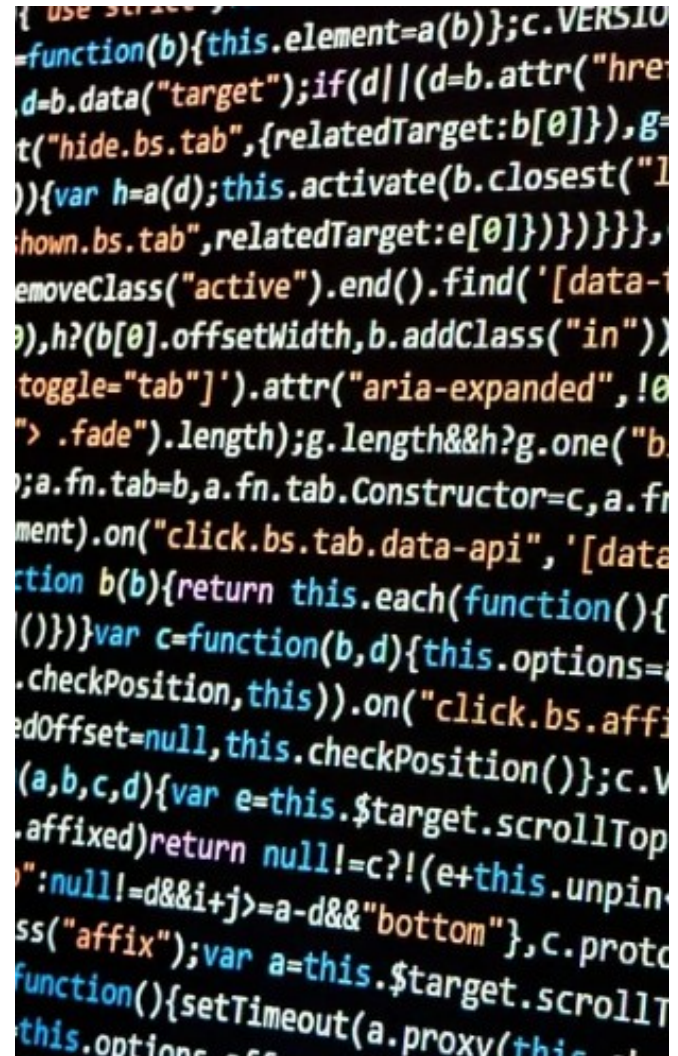

3.4.- Elementos CDATA.

- El formato es el siguiente:

<! [CDATA [texto no procesable...]]>

```
<?xml version="1.0"?>
<documento>
  <título>Prueba</título>
  <ejemplo>
    <![CDATA[
      En HTML la negrita se escribe: <strong>
    ]]>
  </ejemplo>
</documento>
```

- Otro uso de CDATA es colocar dentro de este elemento código de lenguajes de scripts como Javascript para que no sean interpretados como parte de XML.



UT 04.01- El Lenguaje XML. Conceptos Básicos.

3.- Contenido del cuerpo de un documento XML.

3.5.- Espacios de nombres (I).

- Varios documentos XML pueden combinarse, pudiendo coincidir el nombre de algunos elementos.
- **Ejemplo.** Supongamos que tenemos que combinar el contenido de los dos ficheros XML siguientes:

ejemplo1.xml

```
<?xml version="1.0"?>
<carta>
  <palo>Corazones</palo>
  <numero>7</numero>
</carta>
```

ejemplo2.xml

```
<?xml version="1.0"?>
<carta>
  <carnes>
    <carne precio="12.95">Filete de Ternera</carne/>
    <carne precio="13.60">Solomillo a la Pimienta</carne/>
  </carnes>
  <pescados>
    <pescado precio="16.20">Lenguado al Horno</pescado/>
    <pescado precio="15.85">Merluza en Salsa Verde</pescado/>
  </pescados>
</carta>
```

- Si se incluyen ambos elementos <carta> en un documento XML, **se origina un conflicto de nombres.**
- Para resolverlo, se pueden utilizar **espacios de nombres (XML Namespaces).**

UT 04.01- El Lenguaje XML. Conceptos Básicos.

3.- Contenido del cuerpo de un documento XML.

3.5.- Espacios de nombres (II).

- Para definir un espacio de nombres se utiliza la siguiente sintaxis:

xmlns:prefijo="URI"

Los URI especificados en un documento XML no tienen porqué tener contenido, su función es ser únicos. No obstante, en un URI se puede mostrar información si se considera oportuno.

- Por tanto, podríamos hacer lo siguiente:

ejemplo3.xml

```
<?xml version="1.0"?>
<e1:fusion xmlns:e1="ejemplo1"
           xmlns:e2="ejemplo2">

  <e1:carta>
    <e1:palo>Corazones</e1:palo>
    <e1:numero>7</e1:numero>
  </e1:carta>
  <e2:carta>
    <e2:carnes>
      <e2:carne precio="12.95">Filete de Ternera<carne/>
      <e2:carne precio="13.60">Solomillo a la Pimienta<carne/>
    </e2:carnes>
    <e2:pescados>
      <e2:pescado precio="16.20">Lenguado al Horno<pescado/>
      <e2:pescado precio="15.85">Merluza en Salsa Verde<pescado/>
    </e2:pescados>
  </e2:carta>
</e1:fusion>
```

UT 04.01- El Lenguaje XML. Conceptos Básicos.

4.- Documentos bien formados.

4.1.- Definición. Reglas para generar documentos bien formados.



- Un documento XML bien formado **es aquel que cumple con todas las reglas sintácticas definidas para XML.**
- Los procesadores XML pueden rechazar cualquier documento que no esté bien formado. **Un documento XML válido es el que está bien formado.**

Reglas de buena formación.

- Los documentos han de seguir una estructura estrictamente jerárquica en lo que respecta a las etiquetas que delimitan sus elementos.
- Los documentos XML requieren **un solo elemento raíz** en el que todos los demás están incluidos. **La jerarquía de elementos sólo puede tener un elemento inicial.**
- Los elementos no vacíos tienen siempre una etiqueta inicial y una etiqueta final. En su notación abreviada, deben incluir el símbolo “/”, justo antes del cierre de la etiqueta (antes del “>”).
- Los valores de atributos en XML siempre deben estar encerrados entre comillas simples o dobles.
- **XML es sensible a mayúsculas y minúsculas.**
- La especificación XML 1.0 permite el uso de espacios en blanco para tabular el código y **hacer más legible el código. Estos espacios suelen ser tabuladores.**

UT 04.01- El Lenguaje XML. Conceptos Básicos.

4.- Documentos bien formados.

4.2.- Reglas para los nombres y el contenidos.

Reglas para los nombres de marcas, elementos y atributos.

- **Son case sensitive.** Es decir, se distingue entre mayúsculas y minúsculas.
- Pueden contener **letras minúsculas, mayúsculas, números, puntos, guiones medios y guiones bajos.**
- Pueden contener el carácter dos puntos ":". No obstante, **su uso se reserva para cuando se definan espacios de nombres.**
- **El primer carácter tiene que ser una letra o un guion bajo.**
- Detrás del nombre de una etiqueta se permite escribir un espacio en blanco o un salto de línea.

Caracteres que no pueden aparecer dentro del contenido.

- Dentro del contenido de los elementos no pueden aparecer los siguientes caracteres y cadenas, ya que tienen funciones especiales.
 - “<” (comienzo de una etiqueta).
 - “&” (comienzo de una referencia a una entidad).
 - “]]>” (compatibilidad con SGML).

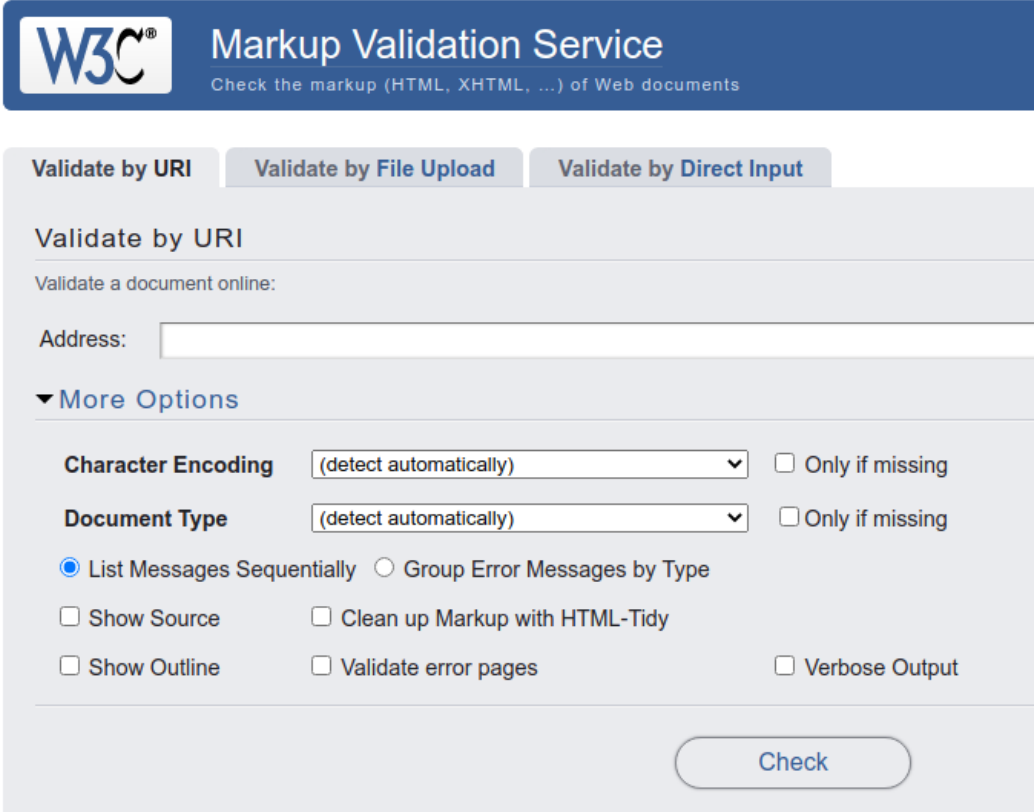


UT 04.01- El Lenguaje XML. Conceptos Básicos.

5.- Herramientas de validación de documentos XML.

- Existen multitud de aplicaciones disponibles para comprobar que un documento XML está bien formado.
- La primera posibilidad es abrir el fichero con un **navegador de Internet. Si está bien formado podremos ver el árbol asociado al documento.**
- Otra posibilidad es utilizar analizadores on line. El consorcio W3C tiene un validador propio accesible en la dirección:

<http://validator.w3.org/>



The screenshot shows the W3C Markup Validation Service interface. At the top, there is a blue header with the W3C logo and the text "Markup Validation Service" and "Check the markup (HTML, XHTML, ...) of Web documents". Below the header, there are three tabs: "Validate by URI", "Validate by File Upload", and "Validate by Direct Input". The "Validate by URI" tab is selected. Under this tab, there is a section "Validate by URI" with the text "Validate a document online:". Below this, there is a text input field labeled "Address:". Below the input field, there is a section "More Options" with several checkboxes and dropdown menus. The "Character Encoding" dropdown is set to "(detect automatically)". The "Document Type" dropdown is also set to "(detect automatically)". There are checkboxes for "Only if missing" next to both dropdowns. Below these, there are two radio buttons: "List Messages Sequentially" (selected) and "Group Error Messages by Type". At the bottom, there are four checkboxes: "Show Source", "Clean up Markup with HTML-Tidy", "Show Outline", "Validate error pages", and "Verbose Output". A "Check" button is located at the bottom right of the form.