

La información y los lenguajes de marcado

Baltasar Fernández Manjón



Dpto. de Sistemas Informáticos y Programación,
Universidad Complutense de Madrid
Avda. Complutense s/n, 28040, Madrid, Spain,

<http://bogart.sip.ucm.es/~balta>

Basado en la conferencia "La información, los lenguajes de marcado y su aplicación a la interacción hombre-computadora" dentro del curso de verano "Interacción Persona-Computador" de la Universidad de Castilla-La Mancha, Puertollano 16 y 17 de septiembre de 1999.

La información, los lenguajes de marcado y aplicaciones

Baltasar Fernández Manjón

La información y los lenguajes de marcado

1. Motivación: World Wide Web
2. Lenguajes de marcado
3. Lenguaje de marcado generalizado estándar (SGML)
4. Estándares relacionados
5. Algunas aplicaciones

¿Exito de Internet y del World Wide Web?

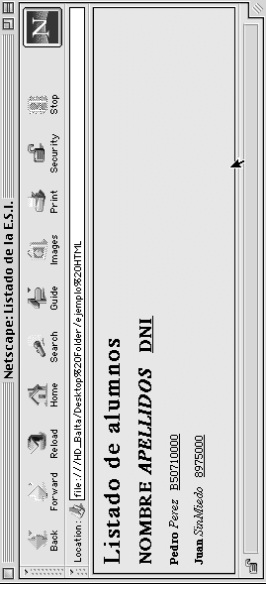
- Web: un nuevo medio de comunicación de masas
 - Revolución en la distribución, acceso e interacción con la información electrónica
 - Información hipermedia: multimedia + hipertexto
- World Wide Web
 - Conjunto de protocolos de comunicación de datos
 - HTTP, FTP, Gopher, etc
 - Sistema de nombrado y localización de recursos (URLs)
 - Formato común de intercambio de datos
 - Lenguaje de marcado de hipertexto (HTML)
- Utilización de un lenguaje de marcado (HTML)
 - Compartición de información entre sistemas heterogeneos
 - Enlazamiento de esa información

Lenguaje de marcado de hipertexto (HTML)

- Separación entre representación y procesamiento
 - Codificación del contenido y estructura de un documento mediante un lenguaje estándar de marcas
 - El navegador realiza y decide la presentación de la información mediante el procesamiento de dichas marcas
- Limitaciones
 - Orientado a la presentación de información
 - No es extensible ni adaptable (conjunto fijo de marcas)
- HTML es una aplicación del Lenguaje de marcado generalizado estándar (SGML)
- Futuro HTML
 - XML - Lenguaje de marcado extensible

Ejemplo HTML

```
<HTML>
<HEAD>
<TITLE>Listado de la E.S.I. </TITLE>
</HEAD>
<BODY>
<H1> Listado de alumnos </H1>
<H2> <B>NOMBRE</B> <I>APELLIDOS</I> <U>DNI</U> </H2>
<P> <B>Pedro</B> <I>Perez</I> <U>B50710000</U>
<P> <B>Juan</B> <I>SinMiedo</I> <U>8975000</U>
</BODY>
</HTML>
```



©Baltasar Fernández Manjón 5

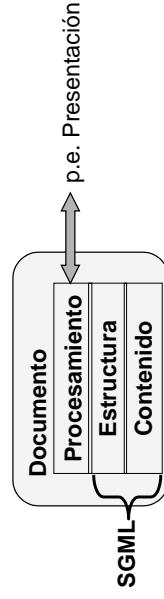
Lenguajes de marcado generalizado

- Técnica descriptiva para representar documentos electrónicos simplificando su procesamiento automático
- Marcado generalizado
 - Descripción, mediante marcas, de la estructura y otras características del documento
 - Metainformación: Estructura, organización, propósito
 - Utilización de la misma información con diferentes finalidades
 - Técnica rigurosa para permitir procesamiento automático
- Documento = Contenido + Marcas
- Maximiza la usabilidad de la información
 - Accesibilidad, mantenibilidad, reutilización
 - Distintos procesamientos, independencia de plataforma

©Baltasar Fernández Manjón 6

Lenguaje de marcado generalizado estándar (SGML)

- Metalenguaje estándar para la definición de lenguajes de marcado (ISO 8879, 1986)
- Método formal y general para diseñar clases de documentos estructurados para cada necesidad concreta
 - Estructuración del documento dependiente del contenido, significado o utilización de los datos
 - Especifica modelos jerárquicos de documentos
 - Características y composición de cada elemento del documento
 - Sintaxis de marcado para definir los elementos
 - Marcas que delimitan cada elemento y atributos aplicables



©Baltasar Fernández Manjón 7

Definición del tipo de documento (DTD)

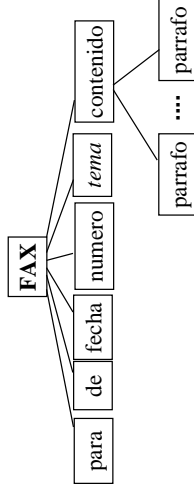
- Cada clase genérica de documentos se representa mediante su Definición de Tipo de Documento (DTD)
 - Conjunto de reglas que especifican cual es la estructura lógica del documento como una jerarquía de elementos marcados (árbol)
 - Orden de elementos, su obligatoriedad, número de ocurrencia, marcas delimitadoras, tipo de contenido
 - Atributos
 - Información complementaria sobre los elementos
- Validación automática de documentos
 - Dado un documento se puede verificar automáticamente si cumple el conjunto de reglas especificado en su DTD
- SGML, no se ocupa de la presentación de los documentos
 - La presentación es sólo un tipo más de procesamiento

©Baltasar Fernández Manjón 8

Ejemplo: definición de tipo de documentos FAX

Fax DTD

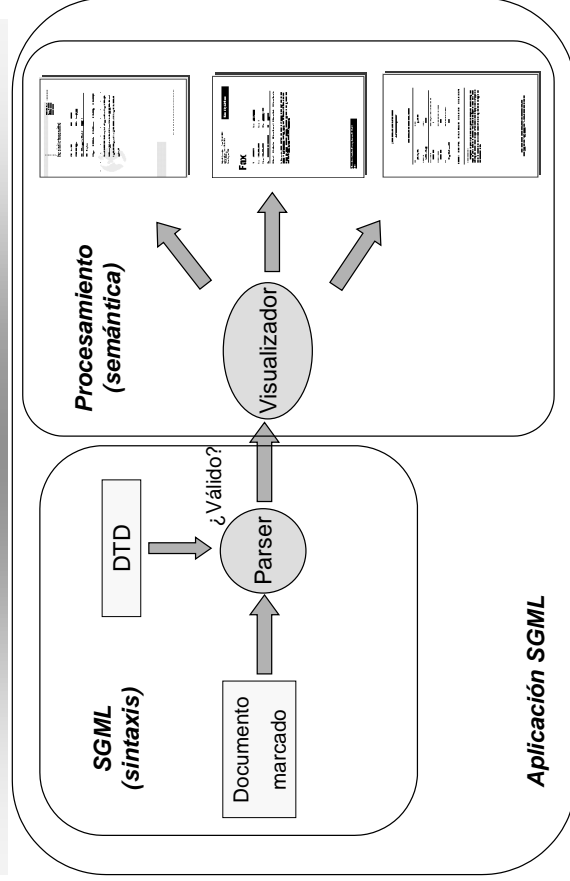
```
<!DOCTYPE fax [
<ELEMENT fax -- (para, de, fecha, numero, tema?, contenido)>
<ELEMENT para -- (#PCDATA)>
<ELEMENT de -- (#PCDATA)>
<ELEMENT fecha -- (#PCDATA)>
<ELEMENT numero -- (#PCDATA)>
<ELEMENT tema -- (#PCDATA)>
<ELEMENT contenido -- (parrafo+)>
<ELEMENT parrafo -- (#PCDATA)>
]>
```



FAX: documento marcado

```
<!DOCTYPE fax SYSTEM "fax.dtd">
<fax>
<para> Manuel Ortega </para>
<de> Baltasar Fernández </de>
<fecha> 16/9/99 </fecha>
<numero> 926 29 53 00 </numero>
<tema> Curso Interacción Persona-Computador </tema>
<contenido>
<parrafo> Estimado Manuel , </parrafo>
<parrafo> Te confirmo nuestra llegada el día 16 en el tren
de las 8. </parrafo>
<parrafo> Un saludo </parrafo>
</contenido>
</fax>
```

Procesamiento de documentos: distintas presentaciones



Componentes básicas de SGML

- Elementos
 - Información textual contenida entre marcas
- Atributos
 - Información complementaria sobre los elementos
 - Pares nombre-valor dentro de una etiqueta
 - Ej. <precio moneda="dolar"> 100 </precio>
- Entidades
 - Modularidad de la información
- Instrucciones de proceso
 - Información procedimental que se incluye en el documento marcado
 - Está destinada a la aplicación que procesa el documento

Utilización de SGML

- Grandes corporaciones que necesitan gestionar ingentes cantidades de documentos durante largos periodos
 - Aviónica
 - Defensa
 - Editoriales
- Se han definido DTD estándares para simplificar la compartición de información
 - Editoriales: **ISO 12083 para publicación de libros y revistas**
 - Software: **DocBook para documentación técnica**
- Alto coste y dificultad de implementación
- Mucha experiencia y probada utilidad

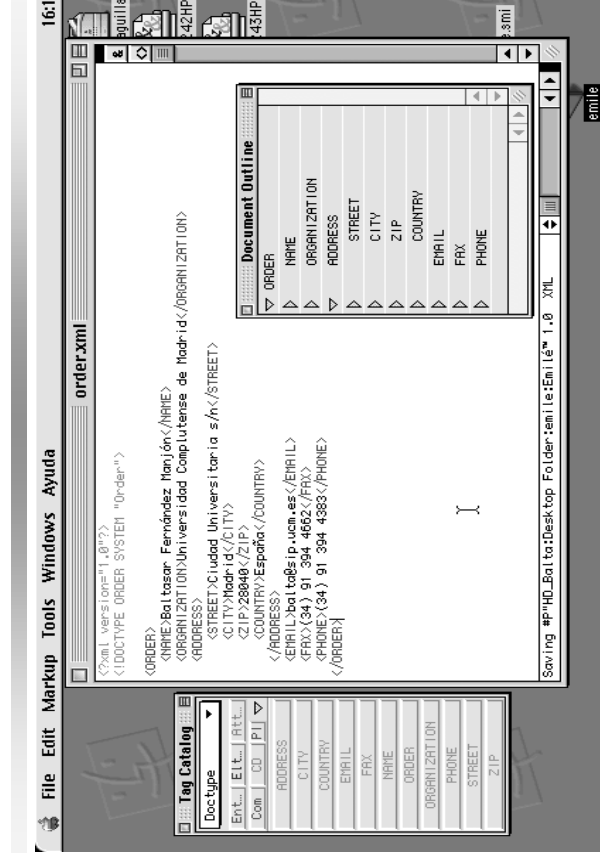
Estándares relacionados con SGML

- Lenguaje de especificación y semántica de estilo de los documentos (DSSSL)
 - Creación de hojas de estilo
 - Permite especificar transformaciones de documentos
- Lenguaje de descripción de aplicaciones hipermedia (HyTime)
 - Mejora la capacidad de representar, direccionar y enlazar información estática y dinámica en cualquier formato o medio
- Lenguajes de propósito específico
 - Lenguaje de marcado de vectores (VML)
 - Lenguaje de marcado para química (CML)

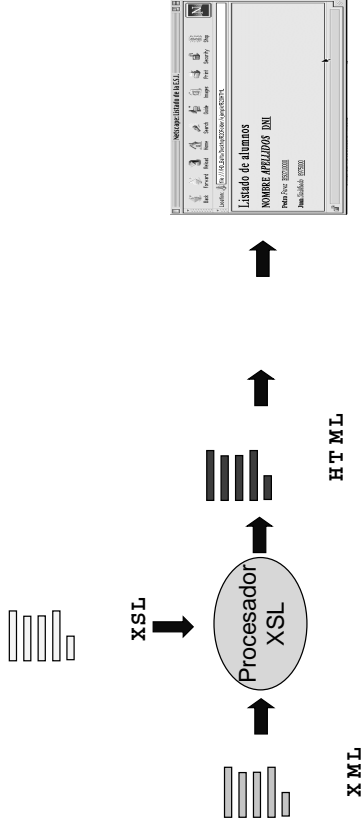
Lenguaje de marcado extensible (XML)

- Lenguaje de marcado extensible (XML)
 - Subconjunto de SGML optimizado para la red
 - Desarrollo más simple de aplicaciones
- Hojas de estilo en cascada (CSS1 y CSS2)
 - Información de estilo para documentos HTML o XML
- Lenguaje de estilo extensible (XSL) y de transformación (XSLT)
 - Basado en DSSSL y CSS
- Otros estándares
 - Xpath define un lenguaje de expresiones para seleccionar elementos de un documento XML (se usa en XSLT)
 - XLink protocolo de gestión de enlace entre documentos XML
 - Xpointer

Editores SGML/XML



Procesamiento documento XML para obtener HTML



17

18

Dos tipos de aplicaciones de los lenguajes de marcado

- Aplicaciones orientadas a presentación y publicación de información
 - Útil con navegadores y editores
 - Normalmente se utiliza para datos que van a ser procesados o presentados a otras personas
- Software de intermediación o de integración orientado al tratamiento de mensajes
 - Útil para el intercambio de datos entre programas con independencia de la plataforma o software específico
 - Definición de configuración de aplicaciones software

Algunas aplicaciones

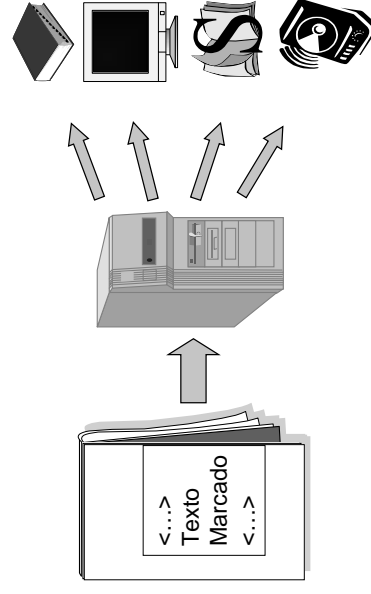
- Accesibilidad de la información
- Personalización y adaptación de contenidos
- Búsqueda “inteligente” de información
- Extracción de información
- Comercio electrónico

19

Accesibilidad de la información

- La separación entre contenido, estructura y presentación permite distintas presentaciones para usuarios discapacitados

- ICADD (International Committee for Accessible Document Design)
- Con un único contenido sólo hay que modificar los procesos de presentación



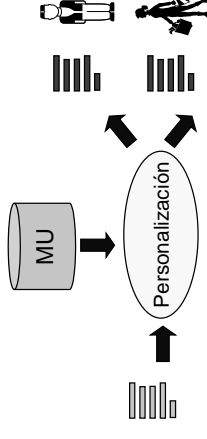
20

Accesibilidad y configuración de documentos

- Documentos interactivos y configurables por el usuario
- Generación automática de ayudas navegacionales a partir de la estructura jerárquica de los documentos
 - Tablas de contenidos navegables
 - Mapas de contenidos
- Distintas presentaciones de un mismo documento
 - Distintas “vistas”
 - Documento con anotaciones
 - ◆ Tres vistas: Sólo documento, sólo anotaciones, ambos contenidos
 - Presentaciones parciales
 - Documento multilingüe
 - ◆ Sólo la parte del idioma seleccionado
 - Documento multiproducto
 - ◆ Sólo las partes comunes y las específicas de ese producto

Personalización de contenidos

- Personalización automática de contenidos y/o interacción a cada usuario concreto
 - Adaptación a las características del usuario
 - Intereses
 - Nivel de conocimiento
 - Preferencias
- Aplicación de técnicas de modelado de usuario (MU)
 - El sistema adquiere y mantiene la información que necesita sobre cada usuario para realizar la adaptación



Búsqueda “inteligente” de información

- Se dispone de grandes cantidades de información en formato electrónico
 - Intranets y Web
 - Búsqueda de información
 - Mediante palabras claves dentro de todo el documento
 - Navegación a través de los resultados (Browsing)
 - Problemas
 - Se recuperan muchos documentos que hay que revisar manualmente
 - Muchos de los documentos recuperados no son relevantes
- Con información marcada las búsquedas pueden ser más precisas ya que se pueden hacer sólo dentro de las partes significativas de un documento
 - Se puede especificar el contexto de la búsqueda
 - Ejemplo: ¿artículos escritos por Bill Gates?
 - ◆ Apariciones sólo dentro de elementos marcados como autor

Búsqueda, extracción y filtrado de información

- El marcado permite integrar datos y texto en los documentos sin perder información
- Actualmente
 - Agentes de compra (shopbots)
 - BargainFinder: Encuentra los precios de CD en diversas tiendas Web
 - ◆ Búsqueda muy simple
 - Agentes activos de búsqueda de información
- La disposición de información marcada permitirá la mejora y sofisticación de estas técnicas
 - Ej. Búsqueda comparativa de productos en distintos países
 - ◆ $\langle \text{precio moneda} = \text{"dolar"} \rangle > 100 \langle / \text{precio} \rangle >$
 - ◆ $\langle \text{precio moneda} = \text{"peseta"} \rangle > 15000 \langle / \text{precio} \rangle >$

Comercio electrónico

- Mayor disponibilidad de servicios y mayor simplicidad de uso
- El mercado facilita el intercambio electrónico de datos
 - Se puede expresar información compleja
 - Es fácil realizar conversiones entre distintos formatos
 - Permite integrar bases de datos y aplicaciones Web
- Se puede separar los datos de sus instrucciones de procesado
- Potenciado por lenguajes de programación multiplataforma
 - Java

Referencias

- Campo muy activo y de rápida evolución
 - Exceso de información
 - Gran cantidad de estándares, recomendaciones, etc
- World Wide Web
 - <http://www.w3.org/>
- SGML
 - <http://www.oasis-open.org/>
- XML
 - <http://xml.org/>