DAI: Mòdul 4
XML

Xavier Noguero
Carles Mateu http://carlesm.com
Ciències de la Computació i Intel·ligència Artificial
Universitat de Lleida

Inconvenients de HTML

- La informació de format i estructura està barrejada.
- La varietat de models per expressar una marca.
- La barreja del format de presentació i la informació de contingut.

Inconvenients de HTML

- La informació de format i estructura està barrejada.

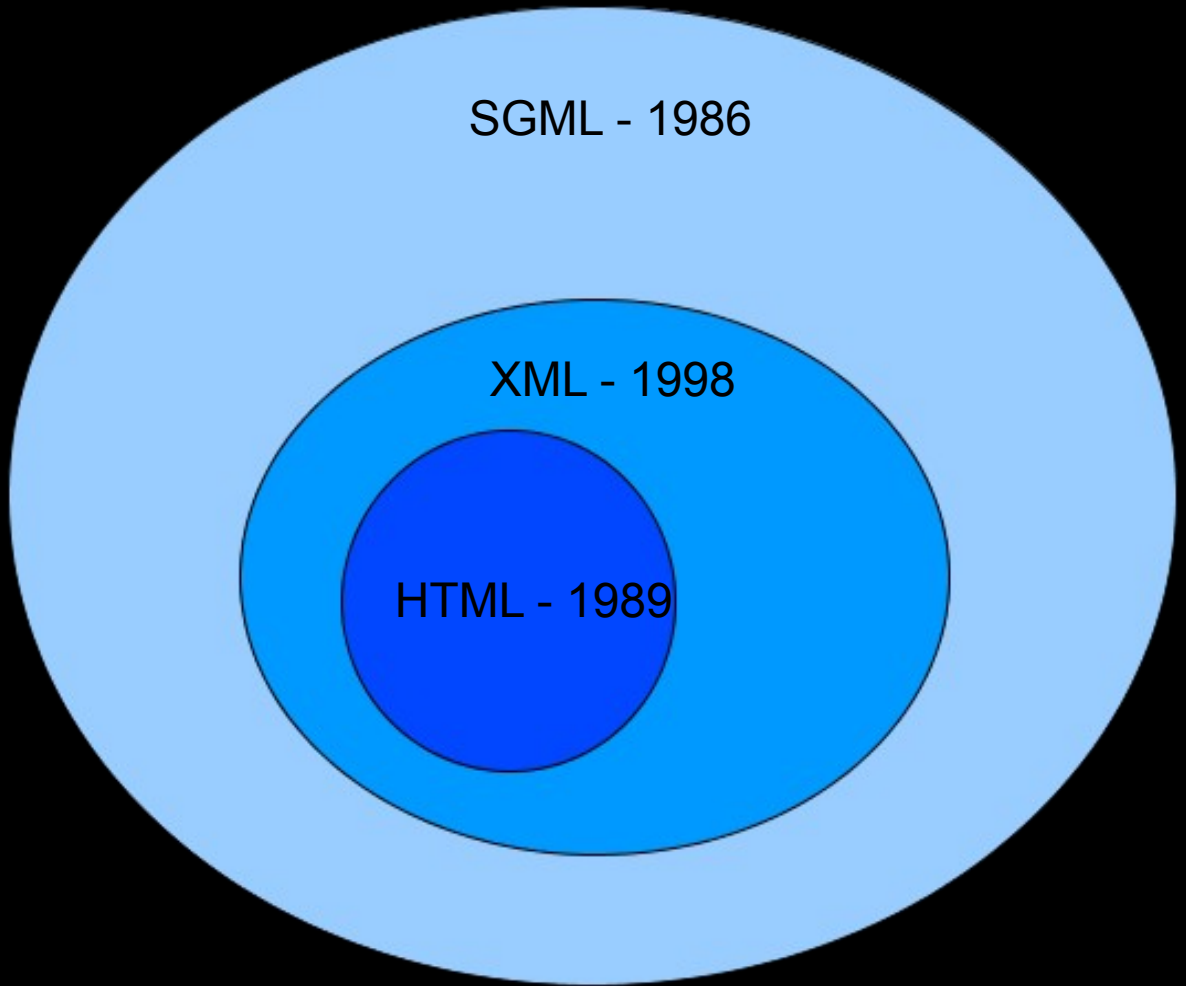
  Solució: fulls d'estil en cascada (CSS)
- La varietat de models per expressar una marca.

  Solució: XHTML
- La barreja del format de presentació i la informació de contingut.

  Solució: XML

Extensible Markup Language (XML)

L' XML (‘llenguatge d’etiquetatge extensible’) és, segons el consorci web W3C, el format universal per a crear documents estructurats i intercanviar dades a la web.



L'objectiu era crear un SGML més senzill, per tal què la indústria trobés rendible invertir en la creació d’eines per al seu tractament.

Comparaţiva HTML - XML

HTML...

```
<table>
<tr>
<td>Hamlet, Príncep de Dinamarca</td>
<td>William Shakespeare</td>
<td>84-239-0027-4</td>
<td>1938</td>
</tr>
</table>
```

XML...

```
<llibre>
<autor>
<cognom>Shakespeare</cognom>
<nom>William</nom>
</autor>
<titol>
Hamlet, Príncep de Dinamarca
</titol>
<isbn>84-239-0027-4</isbn>
<any>1938</any>
</llibre>
```

XML - Finalitats del format

- Ha de ser directament utilitzable sobre Internet.
- Ha de suportar un gran ventall d'usos.
- Ha de ser compatible amb SGML.
- Permetre escriure programes per processar fitxers XML's fàcilment.
- Els fitxers han de ser llegibles i entenedors per als humans.

XML - Finalitats del format

- El disseny XML s'hauria de poder preparar ràpidament.
- El disseny d'XML serà formal i concís.
- Els documents XML han de ser fàcils de crear.
- L'elegància dels tags XML, té poca importància.

Com és un fitxer XML?

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE SYSTEM "exemple.dtd">


<!-- Aquí comencen les dades XML -->
<llibres>
<llibre>
<autor>
<cognom>Shakespeare</cognom>
<nom>William</nom>
</autor>
<titol>Hamlet, Príncep de Dinamarca</titol>
<isbn>84-239-0027-4</isbn>
<any>1938</any>
</llibre>
...
</llibres>
```

Capçalera

Instància del document

Terminologia XML - Elements

- Un document XML està format per un o més elements.
- Cada element representa un component lògic del document.
- Per a descriure els elements també s'utilitzen etiquetes:

```
<títol>Hamlet, Príncep de Dinamarca</títol>
```

Terminologia XML - Elements

- **Principi: <nom_tipus_element [atributs]>**
- **Final: </nom_tipus_element>**

```
<títol>Hamlet, Príncep de Dinamarca</títol>
```

- **Elements buits:<nom_tipus_element/>**

```
<br/> o <img src="paisatge.gif"/>
```

Terminologia XML - Atributs

- Descriuen propietats dels elements.
- Donen informació addicional de l'element.
- Es representen mitjançant parells nom=valor.
- El valor ha d'anar entre cometes simples o dobles.
- Exemple:

```
<títol idioma="català">Hamlet, Príncep de Dinamarca</títol>
```

Terminologia XML - Atributs

Elements versus atributs. Dos exemples:

```
<llibre isbn="00-000-0000-0" titol="titol del llibre"/>
```

Model 1

```
<llibre>
<isbn>00-000-0000-0</isbn>
<titol>titol del llibre</titol>
</llibre>
```

Model 2

Tots dos tenen la mateixa informació
Però n'hi ha un de més adient

Terminologia XML - Atributs

Alguns inconvenients dels atributs:

- No poden tenir múltiples valors.
- No poden contenir estructures d'arbre.
- No es poden expandir en futurs canvis.

Els elements sí, per tant, evitem l'abús d'atributs!

```
<llibre>
<isbn>00-000-0000-0</isbn>
<titol>titol del llibre</titol>
</llibre>
```

Terminologia XML - Entitats

Entitat	Caràcter reservat
<	<
>	>
&	&
"	"
'	'

N'hi ha molt poques comparat amb HTML, etc.

Terminologia XML -Seccions CDATA

- Serveixen per a "escapar" text, de manera que no s'interpreti com a marcat. Sintaxi:

```
<![CDATA[text que volem escapar]]>
```

- Exemple:

```
<script>
<![CDATA[
function suma(a,b)
{
return a+b;
}
]]>
</script>
```

Terminologia XML – Instruccions de procès

- Especifiquen informació per a les aplicacions.
- Sintaxi:

```
<?instruccio atributs?>
```

- Exemple:

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

- Indica que s'ha creat seguint l'especificació 1.0

Terminologia XML – Instruccions de procès

- **Especifiquen informació per a les aplicacions.**
- **Sintaxi:**

```
<?instruccio atributs?>
```

- **Exemple:**

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

- **Indica que la codificació de caràcters és la ISO-8859-1.**

Sintaxi XML – Normes sintàctiques

- Un document és un document XML ben format si obeeix les normes de sintaxi del llenguatge XML.

- Cap document que no estigui ben format és un document XML.

Sintaxi XML – Normes sintàctiques

Els documents XML han de tenir un element arrel únic.

Correcte:

```
<llibres>
<llibre><titol>The world is flat</titol></llibre>
<llibre><titol>Learning xml</titol></llibre>
</llibres>
```

Incorrecte:

```
<llibre><titol>The world is flat</titol></llibre>
<llibre><titol>Learning xml</titol></llibre>
```

Sintaxi XML – Normes sintàctiques

Els documents XML han de tenir una estructura d'etiquetes jerarquitzada.

Correcte:

```
<LLIBRE><TITOL> </TITOL></LLIBRE>
```

Incorrecte:

```
<LLIBRE><TITOL> </LLIBRE></TITOL>
```

Sintaxi XML – Normes sintàctiques

- Es distingeix entre majúscules i minúscules.

Correcte:

```
<titol>The world is flat</titol>
```

Incorrecte:

```
<titol>The world is flat</TITOL>
```

Sintaxi XML – Normes sintàctiques

- Els espais en blanc són irrellevants (s'eliminen o es normalitzen).

- Per tant, són equivalents:

```
<titol> The world is flat </titol>
```

```
<titol>The world is flat</titol>
```

Sintaxi XML – Normes sintàctiques

- Un document és un document XML ben format si s'ajusta a la sintaxi del llenguatge XML. Cap document que no estigui ben format és un document XML.
- Requisits:
- Els documents XML han de tenir un únic element pare.
- Han de tenir una estructura d'etiquetes jerarquitzada.
- Es distingeix entre minúscules i majúscules als elements.
- Els elements han de tenir una etiqueta de tancament.
- Els atributs XML han d'anar entre cometes.

Espais de noms (o namespaces)

- Suposem aquestes propostes XML per a representar dades sobre llibres i autors:

```
<llibre>
<titol>The world is flat</titol>
<isbn>00-000-0000-0</isbn>
</llibre>
```

```
<autor>
<nom>Thomas L. Friedman</nom>
<titol>Història de l'Art</titol>
</autor>
```

Espais de noms (o namespaces)

- Suposem aquestes propostes XML per a representar dades sobre llibres i autors:

```
<llibre>
<titol>The world is flat</titol>
<isbn>00-000-0000-0</isbn>
</llibre>
```

```
<autor>
<nom>Thomas L. Friedman</nom>
<titol>Història de l'Art</titol>
</autor>
```

I es decideix aquesta proposta conjunta:

```
<llibre>
<titol>The world is flat</titol>
<isbn>00-000-0000-0</isbn>
<nom>Thomas L. Friedman</nom>
<titol>Història de l'Art</titol>
</llibre>
```

Espais de noms (o namespaces)

- Suposem aquestes propostes XML per a representar dades sobre llibres i autors:

```
<llibre>
<titol>The world is flat</titol>
<isbn>00-000-0000-0</isbn>
</llibre>
```

```
<autor>
<nom>Thomas L. Friedman</nom>
<titol>Història de l'Art</titol>
</autor>
```

I es decideix aquesta proposta conjunta:

```
<llibre>
<titol>The world is flat</titol>
<isbn>00-000-0000-0</isbn>
<nom>Thomas L. Friedman</nom>
<titol>Història de l'Art</titol>
</llibre>
```

Ambiguitat i Col·lisió!

Espais de noms (o namespaces)

- Sol·lució de XML: mantenim l'etiqueta i hi afegim al davant un prefix que descrigui l'àmbit.
- Cadascun d'aquests àmbits és un espai de noms.
- Declaració d'un namespace:

```
xmlns:prefix = "uri_espai_de_noms"
```

- On la sintaxi d'una URI és:

```
protocol://hostname/path[#fragment]
```

Espais de noms (o namespaces)

- **Una sol·lució sense ambigüitats:**

```
<llibre xmlns:infoLlibre="uri_espainoms_llibres"
xmlns:infoAutors="uri_espainoms_autors">
<infoLlibre:titol>The world is flat</titol>
<infoLlibre:isbn>00-000-0000-0</isbn>
<infoAutors:nom>Thomas L. Friedman</nom>
<infoAutors:titol>Història de l'Art</titol>
</llibre>
```

Ara queda clar quin és el títol del llibre i quina titulació té l'autor

Més d'XML

- XML és un mètode per a posar dades estructurades en text.
- XML sembla HTML però no és HTML.
- XML és text, però no està fet per a ser llegit.
- XML engloba una família de tecnologies.
- XML és molt extensible, però això no és cap problema...
- XML és nou, però no tant nou.
- XML no necessita llicència, és independent de la plataforma i està ben suportat per moltes eines.

Usos de XML

- Representar i transportar informació estructurada, com la que es pot emmagatzemar en una base de dades.
- Representar informació que s'hagi d'ensenyar a persones (en això es basa XHTML).
- Diversos protocols l'usen per codificar les dades que s'intercanvien (WebDAV, serveis web, XMPP, entre molts d'altres).

Resum de característiques de XML
XML

- Els elements...
  - són decidits per l'usuari.
  - poden tenir atributs.
  - sense dades es poden tancar al final de l'etiqueta.
  - s'han de tancar sempre!
- Codificació d'alguns caràcters.
- Marques, distingeixen el cas.

Document Type Definition (DTD)

- Un documents DTD és el conjunt de restriccions que s'han de seguir per a crear una representació d'un document XML d'un tipus determinat.
- La seva funció principal és permetre de validar de manera automàtica si un document XML compleix aquestes restriccions o no.
- Per especificar aquestes normes, s'utilitza la notació de declaració de marcatge, descrita a l'especificació XML.

Document Type Definition (DTD)

Que podem indicar amb declaracions de marcatge?

- Els elements i atributs vàlids dins d'un document XML.
- Els elements que es poden utilitzar dins d'altres elements.
- Els elements i atributs opcionals.

Document Type Definition (DTD)

DTD:

```
<!ELEMENT note (to, from, heading, body)>
<!ELEMENT to (#PCDATA)>
<!ELEMENT from (#PCDATA)>
<!ELEMENT heading (#PCDATA)>
<!ELEMENT body (#PCDATA)>
```

possible XML:

```
<note>
<to>Estudiants de DAI</to>
<from>Tim Berners-Lee</from>
<heading>Salutacions</heading>
<body>Salutacions cordials!</body>
</note>
```

- **Un document XML és vàlid si el seu contingut respecta les regles del seu document DTD associat.**

Document Type Definition (DTD)

- **No confondre document XML ben format amb document xml vàlid**

  Tots els documents xml han d'estar ben formats.

  Un document xml, serà vàlid només si compleix les normes del DTD que té associat.
- Si un document xml no té cap DTD associat, no podem dir res sobre la seva validesa.

Document Type Definition (DTD)

Els DTD tenen bàsicament dos problemes:

- S'especifiquen en un llenguatge totalment diferent d'XML.
- Tenen molt poca riquesa per a expressar tipus de dades.

Per a solucionar aquests problemes, es va crear un altre mecanisme alternatiu per a validar XML's: l'especificació XML Schema.

XML SCHEMA

L'especificació **XML Schema** és una notació per a escriure DTDs alternatius a les declaracions de marcatge.

Bàsicament, aporta tres avantatges:

- **És notació XML:** això fa que les mateixes eines que s'utilitzen per a analitzar XML es puguin utilitzar per a analitzar esquemes (schema) XML.
- **Suporta diferents tipus de dades:** ofereix un conjunt de tipus de dades ja definits (ex: tipus de calendari, de moneda, numèrics, etc.) i això permet una validació molt més eficaç dels documents XML.
- **És extensible:** permet noves definicions de tipus de dades.

Núvol d'etiquetes

XML

XLink

XSL

XPath

RDF

SVG

XPointer

XSL-fo

RSS

Namespaces

XSLT

XML Query

DOM

RDF

CSS

SMIL

XML Encription

SAX

XForms

DTD

Schema

XML Signature

XML Base

SOAP

Web Ontology

Semantic Web

XInclude

WSDL

WAP

Bibliografia XML

- Especificació XML – W3C:

http://www.w3.org/TR/REC-xml

- Tutorial i exemples pràctics XML:

http://www.w3schools.com/xml/default.asp

- Namespaces:

http://www.xml.com/pub/a/1999/01/namespaces.html

- Tutorial DTD:

http://www.w3schools.com/dtd/default.asp

- Especificació XML schema:

http://www.w3.org/XML/Schema