INTRODUCCIÓ ALS LLENGUATGES DE MARQUES

lloc: CIFP Francesc de Borja Moll
Curs: Llenguatges de marques i sistemes de gestió d'informació
Llibre: INTRODUCCIÓ ALS LLENGUATGES DE MARQUES
Imprès per: Joan Llompart Socias
Data: divendres, 5 novembre 2021, 10:30

Taula de continguts

1. Introducció als llenguatges de marques

Una de les tasques bàsiques que fan els ordinadors és emmagatzemar la informació que els proporcionem per poder ser processada després. Aquesta informació pot ser de molts de tipus diferents (text, imatges, vídeos, música…) però el més important serà de quina manera l'emmagatzema l'ordinador per poder-la tractar posteriorment de manera eficient per generar més informació.

RECORDATORI:
Una definició senzilla del que és un ordinador podria ser: "Una màquina electrònica que rep i processa dades per convertir-les en informació útil".

RECORDATORI:
Les dades són representacions d'aspectes del món real i se solen recollir per fer càlculs, mostrar-les, organitzar-les, ..., amb l'objectiu que posteriorment algú en pugui fer alguna cosa amb elles, per exemple: prendre decisions, generar noves dades, …

RECORDATORI:
Podríem definir de forma molt general un sistema informàtic com un sistema on les úniques tasques que es desenvolupen consisteixen a emmagatzemar dades per processar-les per mitjà d'un programa que o bé aportarà algun tipus d'informació o bé es faran servir de nou per generar noves dades.

Característiques de les dades

Les característiques més importants de les dades en podem basar en tres aspectes:

- A qui van dirigides.
- La possibilitat de reutilitzar.
- Que es puguin compartir.

    - **A qui van dirigides:**

      Si s'intenta ser una mica més pràctic, es veurà que realment les dades tindran una forma o una altra en funció del destinatari a qui vagin dirigides:

      - Dades destinades a les persones: Aquestes dades hauran de tenir alguna

estructura concreta, amb uns formats determinats, per exemple, hi apareixeran títols, caràcters en negreta, ... Generalment no serà necessari conèixer quin significat tenen les dades, ja que la interpretació es deixarà al lector.

- Dades destinades als programes: Els programes generalment no necessiten que les dades tinguin informació sobre com s'han de representar, simplement serà necessari que siguin fàcilment identificables, que quedi clar de quin tipus són i que hi hagi alguna manera de determinar el que signifiquen per poder-les tractar automàticament.

- **La possibilitat de reutilitzar:**

Moltes vegades, les dades es voldran reutilitzar per poder fer tasques diferents. Un error corrent és emmagatzemar-les en funció d'una tasca concreta, ja que això pot provocar que posteriorment sigui molt més complicat fer-les servir per fer altres tasques. Per tant, és bàsic disposar d'un sistema d'emmagatzematge que permeti aconseguir que les dades puguin ser reutilitzades fàcilment i si pot ser, que puguin ser reutilitzades tant per les persones com pels programes.

- **Compartició de les dades**

En els inicis de la informàtica, els ordinadors generaven i processaven la informació en el mateix lloc. Però l'aparició dels ordinadors personals, l'eclosió de les xarxes i, sobretot, l'èxit d'Internet, ha creat tota una sèrie de problemàtiques que fins al moment no existien:

IMPORTANT:

```
"Les dades generades en un lloc ara poden ser consumides en un
lloc totalment diferent, per exemple, en sistemes operatius tot
alment diferents, en màquines que poden funcionar de maneres mo
lt diverses".
```

Per tant, en un sistema informàtic modern s'ha de tenir en compte aquesta possibilitat a l'hora d'emmagatzemar dades. Hi ha la possibilitat que aquestes dades siguin compartides i, per tant, és necessari emmagatzemar-les de qualque manera per evitar tenir problemes per emprar-les en sistemes diferents.

Emmagatzematge de dades en ordinadors

En l'actual arquitectura dels ordinadors (*Von Neumann*), la informació que s'hi pot emmagatzemar sempre es representa mitjançant uns i zeros (1, 0), és a dir, emprant el sistema binari. Això fa que per representar qualsevol classe de dades (imatges, vídeos, text…) sigui necessari fer algun tipus de procés que converteixi les dades a una representació en format binari.

Tradicionalment en els ordinadors les dades s'organitzen de dues maneres:

- **Dades binàries**

  Emmagatzemar les dades de manera binària és la manera natural d'emmagatzemar dades en ordinadors. Són una tira de bits un darrere l'altre. Les dades en format binari tenen una sèrie de característiques que les fan ideals per als ordinadors:

  - Generalment estan optimitzades per ocupar només l'espai necessari.
  - Els ordinadors les llegeixen fàcilment.
  - Poden tenir estructura.
  - És relativament fàcil afegir-hi metadades (dades que defineixen i descriuen altres dades).
  - Estan disponibles immediatament per fer càlculs numèrics, ja que realment es tracta de nombres. No serà necessari fer cap transformació per poder emprar aquests números en qualsevol càlcul.

  Si un programa vol emprar les dades binàries directament, necessitarà conèixer la mida en bits i, sobretot, conèixer de quina manera s'hi ha emmagatzemat la informació. Per exemple: Per emmagatzemar el nombre 150 només serà necessari convertir aquest valor decimal a la seva representació en binari i emmagatzemar-lo. És trivial comprovar que pot ser emmagatzemat en un sol byte (8 bits):

valor decimal	valor binari
150	10010110

  Un problema és que les dades en format binari estan pensades per ser llegides per màquines, però no per les persones, de manera que són ideals per ser emmagatzemades en màquines, van bé per a la comunicació d'informació entre màquines, però en canvi perquè un humà les pugui fer servir serà necessari tenir un programa específic per llegir-les.

- **Dades de text**

  Per solucionar el problema de recuperar les dades que hi ha en un fitxer, existeix una possibilitat que és fer el més obvi, fer el mateix que han fet les persones durant segles. Els humans en escriure ja estan fent servir una codificació i, per tant, si es fa servir la mateixa codificació, tindrem les dades en un format fàcil d'entendre i perquè es puguin emprar, per tant no hi haurà problemes perquè el codi pugui ser llegit pels programes.

***AMPLIACIÓ*:**

- En fitxers binaris, el component d'informació més petit és el bit.
- En fitxers de text el component més petit és el caràcter.

Els fitxers de text emmagatzemen la informació lletra per lletra d'una manera similar a com ho faria una persona en escriure. Això fa que s'estigui generant una informació que es podrà llegir de la mateixa manera que es llegeix un document de paper. Per a un ordinador no hi ha gaire diferència a l'hora d'emmagatzemar els fitxers de text o els fitxers binaris, ja que els fitxers de text també són tires de bits. La diferència és que en els fitxers de text, els bits estan agrupats d'una manera estàndard i coneguda: un codi de caràcters.

AMPLIACIÓ:
Representar les dades en un ordinador en forma de text implica que per poder representar una paraula qualsevol a l'ordinador, prèviament haurà de ser codificada perquè pugui ser representada en binari (recordem que els ordinadors només poden representar dades en binari).
Aquesta codificació consisteix a determinar una quantitat de bits predefinida per marcar un caràcter, i posteriorment, s'associa un valor numèric a cada un dels caràcters.

1.1. Definició i classificació dels llenguatges de marques

IMPORTANT:

```
Els llenguatges de marques o llenguatges de marcat són aquells que combinen dins un document, la informació (generalment textual) amb marques (o anotacions relatives a l'estructura del text o de la forma de representar-lo).
```

El llenguatge de marques és el que especifica quines seran aquestes marques possibles (etiquetes, elements, etc.) , on s'han de col·locar i el significat que tindrà cadascuna d'elles. A més, la presència de marques intercalades en el contingut fa explícita l'estructura del document o qualsevol informació addicional que es vulgui ressaltar.

Els documents que es creen amb el llenguatge de marques tenen com a avantatge la facilitat de creació i lectura. Això és gràcies al compliment d'estàndards d'emmagatzematge definits i públics, a la incorporació de metadades i a la definició de l'estructura de les dades.

Com que els fitxers de text sempre estan emmagatzemats en algun codi de caràcters conegut (per exemple: ASCII, UTF-8, etc.) s'aconsegueix que puguin ser transportats i llegits en qualsevol plataforma, sistema operatiu o programa que pugui interpretar aquests codis de caràcters. Per tant, els llenguatges de marques s'aprofitaran d'aquesta característica, en estar basats en el format de text. A més, també tindran l'avantatge que podran ser oberts i creats amb els programes d'edició de text estàndard. Des d'editors tan simples com el Bloc de notes dels sistemes Windows o el Gedit de sistemes Unix, fins a editors més complexo, passant per editors especialitzats en XML com Atom, Visual Studio Code, Sublime Text, Adobe Dreamweaver, Oxygen XML Editor, XML Copy Editor, etc, i també existeixen *frameworks* . També permeten definir les dades i la seva estructura de manera que sigui senzill per un programa poder-les interpretar.

Gràcies als avantatges que ofereixen els llenguatges de marques, aquests s'han convertit ràpidament en una de les maneres habituals de representar dades i es poden trobar contínuament en les tasques habituals amb ordinadors:

- L'exponent més popular és Internet –el Web–, que està basat totalment en els llenguatges de marques.
- Molts dels programes d'ordinador que es fan servir habitualment utilitzen en un moment o altre, algun llenguatge de marques per a emmagatzemar les seves dades de configuració o de resultats.

Les marques

Les marques són una sèrie de codis que s'incorporen als documents electrònics per determinar-ne el format, la manera com s'han d'imprimir, l'estructura de les dades, etc. Per tant, són anotacions que s'incorporen a les dades.
Les marques més emprades són les que estan formades per texts descriptius i estan envoltades dels símbols de "més petit" (<) i "més gran" (>) i normalment n'hi sol haver una al principi i una al final:

<marca>....</marca> de forma general

Aparició i evolució dels llenguatges de marques

- La idea del marcat procedeix de l'anglès "*marking up*", terme amb el qual es referien a la tècnica de marcar manuscrits amb llapis de color per fer anotacions com ara la tipografia a emprar en les impremtes. Aquest mateix terme s'ha utilitzat per als documents de text que contenen ordres o anotacions.
- Les possibles anotacions o indicacions incloses en els documents de text han donat lloc a llenguatges (entenent que en realitat són formats de document i no llenguatges en el sentit dels llenguatges de programació d'aplicacions) anomenats llenguatges de marques, llenguatges de marcat o llenguatges d'etiquetes.
- Es considera a Charles Goldfarb com el pare dels llenguatges de marques. Es tracta d'un investigador d'IBM que va proposar idees perquè els documents de text tenguessin la possibilitat d'indicar el seu format. Va contribuir a definir el llenguatge GML d'IBM, el qual va posar les bases del llenguatge SGML (pare de HTML i XML) ideat per Goldfarb.
- A finals dels anys 80 dins el CERN (*Conseil Européen pour la Recherche Nucléaire*) es va crear un llenguatge de marcat pensat per compartir informació usant les xarxes d'ordinadors i, de forma més general, a través d'Internet. Aquest llenguatge es basava en alguns principis de SGML i ho van denominar HTML (*Hyper-text de marques*). L'aparició d'aquest llenguatge va suposar d'alguna manera una revolució en la forma de compartir informació, gràcies principalment a la senzillesa de la seva sintaxi i del programari necessari per a interpretar-lo. En poc temps el llenguatge HTML es va estendre i va començar a créixer de forma a vegades descontrolada i gairebé sempre influenciat per raons merament comercials.
- A mitjans dels anys 90 el consorci W3C (*World Wide Web Consortium*) va començar una iniciativa per intentar dotar la web d'un llenguatge més potent i que pogués donar una estructurar semàntica a aquesta. Per a això es van marcar l'objectiu de crear un nou llenguatge de marques basat en SGML i que fos senzill com HTML. Finalment, l'any 1998, W3C va fer públic un nou estàndard que van denominar XML (*eXtended Markup Language*), més senzill que SGML i més potent que HTML.

Característiques dels llenguatges de marques

Els llenguatges de marques han destacat per una sèrie de característiques que els han convertit en els tipus de llenguatges més emprats en la informàtica actual per emmagatzemar i representar les dades. Entre les característiques més interessants que ofereixen els llenguatges de marques es troben:

- Que es basen en el text pla.
- Que permeten fer servir metadades.
- Que són fàcils d'interpretar i processar.
- Que són fàcils de crear i bastant flexibles per representar dades molt diverses.
- Les aplicacions d'Internet i molts dels programes d'ordinador que es fan servir habitualment, fan servir de qualque manera o altra algun llenguatge de marques.

Avantatges dels llenguatges de marques

- Es poden interpretar directament perquè que fan servir el format de text.
- Són independents de la plataforma, del sistema operatiu o del programa.
- El fet que estiguin basats en format de text fa que siguin fàcils de crear i de modificar perquè només requereixen un simple editor de texts.
- Facilitat de procés: Permeten que el processament de les dades que contenen pugui ser automatitzat de qualque manera, ja que el fitxer conté l'estructura de les dades i això fa que programa pugui interpretar cada una de les dades d'un fitxer de marques per representar-lo o tractar-lo convenientment, ja que mostren l'estructura de les dades que contenen. Posteriorment un programa podrà interpretar gràcies a les marques què és el que significa cada una de les dades del document.

Classificació dels llenguatges de marques

Podem classificar els llenguatges de marques en dos grans grups basats en el seu objectiu:

1. Llenguatges descriptius o semàntics: Orientats a descriure l'estructura de les dades que conté. En aquests llenguatges es descriu quina estructura lògica té el document ignorant de quina manera serà representada en els programes. Només es posen les marques amb l'objectiu de definir les parts que donen estructura al document. L'exemple més important és l'XML.

Exemple fragment document XML:

```
<carta>
   <data>01/11/2020</data>
   <salutacio>Estimat company:</salutacio>
   <contingut> contingut de la carta ...</contingut>
   <firma>Adela Bujosa</firma>
</carta>
```

2. Llenguatges procedimentals i de presentació: Orientats a especificar com s'ha de representar la informació. En aquests llenguatges el que es fa és indicar de quina manera s'ha de fer la presentació de les dades. Ja sigui per mitjà d'informació per al disseny (marcar negretes, títols, ...) o de procediments que ha de fer el programari de representació. L'exemple més popular d'aquests llenguatges és l'HTML però n'hi ha molts més: TeX, Wikitext… En aquests casos els documents ens poden servir per determinar de quina manera es mostrarà el document a qui el llegeixi.

Exemple fragment document HTML:

```
<html>
 <head>
  <title>Exemple senzill</title>
 </head>
 <body>
  <p>Aquest text és un paràgraf.</p>
 </body>
</html>
```

AMPLIACIÓ:
Sistema d'etiquetatge: Tant si el sistema és descriptiu com de presentació, les marques no han estat col·locades de qualsevol manera sinó que s'ha anat seguint un sistema determinat. Sovint les marques envolten el contingut que volem que tingui un significat o que sigui representat d'una manera determinada. No es poden col·locar les marques de qualsevol manera, ja que una de les coses que cal evitar són possibles mal interpretacions.
Per això, a més de definir les marques que s'hi posaran, els llenguatges de marques defineixen unes regles d'ús que especifiquen com han de ser les marques, en quines condicions es permet fer-les servir i a vegades fins i tot què significiquen.

Utilització de llenguatges de marques en entorns web

Una pàgina web és un document electrònic adaptat per a la *World Wide Web* que, normalment, forma part d'un lloc web.

La pàgina web, està composta principalment per informació (només text o mòduls multimèdia) així com per hiperenllaços; a més, pot contenir o associar dades d'estil per especificar com ha de visualitzar-se, i també aplicacions embegudes per fer-la interactiva (Una aplicació embeguda es tracta d'un programa categoritzat dins de la família del software de sistema que està directament integrat en un sistema de hardware i la seva finalitat és controlar màquines o dispositius. Generalment està dissenyat pel hardware particular en el que s'executa i a més compleix una única funció per el que no pot ser utilitzat en altres situacions).

Les pàgines web estan escrites en un llenguatge de marques que proporciona la capacitat de gestionar i inserir hiperenllaços, generalment, HTML.

El contingut de la pàgina pot ser predeterminat (pàgina web estàtica) o generat en el moment de la seva visualització o en sol·licitar-la a un servidor web (pàgina web dinàmica). Pel que fa a l'estructura de les pàgines web, alguns organismes, especialment el W3C, solen establir directives amb la intenció de normalitzar el disseny, per tal de facilitar i simplificar la visualització i interpretació del contingut.

1.2. Organitzacions desenvolupadores

Dins de les organitzacions que s'han encarregat de desenvolupar els llenguatges de marques es troben:

- Organització Internacional per a l'Estandardització (*ISO, International Organization for Standardization*)

Es va formar després de la Segona Guerra Mundial (23 de febrer de 1947) i és l'organisme encarregat de promoure el desenvolupament de normes internacionals de fabricació, comerç i comunicació per a totes les branques industrials a excepció de l'elèctrica i l'electrònica.

La seva funció principal és la de cercar i definir l'estandardització de normes de productes i seguretat per a les empreses o organitzacions en l'àmbit internacional. És una xarxa dels instituts de normes nacionals de 163 països, sobre la base d'un membre per país, amb una Secretaria Central a Ginebra (Suïssa) que coordina el sistema.

Les normes desenvolupades per ISO són voluntàries, ja que és un organisme no governamental i no depèn de cap altre organisme internacional, per tant, no té autoritat per imposar les seves normes a cap país. El contingut dels estàndards està protegit per drets d'autor i per accedir-hi al públic en general ha de comprar cada document. Aquesta organització després de l'èxit que va tenir GML i, després d'un llarg procés, va publicar el 1986 *l'Standard Generalized Markup Languaje* (SGML) amb rang d'estàndard internacional amb el codi ISO 8879.

- W3C (*World Wide Web Consortium*)

El W3C es va crear el 1994 per Tim Berners-Lee al MIT, actual seu central del consorci. Posteriorment es va unir l'abril de 1995, l'INRIA a França, reemplaçat pel ERCIM el 2003 com l'hoste europeu del consorci i la Universitat de Keiō (*Shonan Fujisawa Campus*) al Japó el setembre de 1996 com a hoste asiàtic.

La seva funció principal és tutelar el creixement i organització de la web. El seu primer treball va ser normalitzar el llenguatge HTML, el llenguatge de marques amb què s'escriuen les pàgines web. En créixer l'ús del web, van créixer les pressions per ampliar l'HTML. El W3C va decidir que la solució no era ampliar l'HTML, sinó crear unes regles perquè qualsevol pogués crear llenguatges de marques adequats a les seves necessitats, però mantenint unes estructures i sintaxi comunes que permetessin compatibilitzar i tractar-los amb les mateixes eines. Aquest conjunt de regles és l'XML, la primera versió es va publicar en 1998.

1.3. Etiquetes, elements i atributs dels llenguatges de marques

A l'apartat anterior hem explicat que són els llenguatges de marques, i s'han introduït alguns conceptes que ara es fa necessari explicar en detall:

Hi ha tres termes emprats per tots els llenguatges de marques, que s'utilitzen per descriure les parts d'un document de llenguatges de marques:

- **Elements**: Representen estructures mitjançant les quals s'organitzarà el contingut del document o accions que es desencadenen quan el programa navegador interpreta el document. Consten de l'etiqueta d'inici, l'etiqueta de cap i de tot allò que es troba entre les dues. Alguns elements no tenen contingut. Se'ls denomina elements buits i no han de dur cap etiqueta.
- **Etiqueta o tag**: És un text que va entre el símbol menor que (<) i el símbol més gran que (>). Existeixen etiquetes d'inici (ex. <nom>) i etiquetes de fi (ex. </ nom>).
- **Atribut**: És un parell nom-valor que es troba dins de l'etiqueta d'inici d'un element i indiquen les propietats que poden portar associades els elements.

Exemple amb codi HTML:

```
<html>
  <head>
    <title>Document</title>
  </head>
  <body>
    <h1>Exemple</h1>
  </body>
</html>

On tenim:
- Un element pare <html>, que té 2 elements fills <head>, <body
>. L'element <head> té un element fill <title> sense fills. L'e
lement fill <body> que té un element fill <h1>.
- Les etiquetes que són totes les paraules que estan entre els
símbols < >.
- No hi ha atributs en aquest exemple.
```

Exemple amb codi XML:

```
<adreça>
   <client>
      <nom> Maria </nom>
      <llinatges> Más López </llinatges>
      <carrer> Dels tarongers, 12 </carrer>
      <provincia ciutat="Palma de Mallorca"> Illes Balears</pro
vincia>
      <codi_postal> 07002 </codi_postal>
   </client>
</adreça>
```

On tenim:
- Un element pare <adreça>, que conté 1 element fill <client>. L'element <client> té 5 elements fills: <nom>, <llinatges>, <carrer>, <provincia> i <codi_postal>.
- Les etiquetes que són totes les paraules que estan entre els símbols < >.
- Un atribut, que està dins l'element <provincia> i que és ciutat.