# M4. Llenguatges de marques – UF 1 - XML

INSTITUT PÚBLIC DE SABADELL

Marc Albareda

Eloi Vazquez

# INDEX

- En aquest PPT veurem:
  - Que és un llenguatge de Marques
  - Tipus de llenguatges de Marques
  - Introducció a la codificació de caràcters
  - Introducció XML

# LLENGUATGE DE MARQUES

- Què és un llenguatge de Marques?
  - És una manera de codificar un document de text de manera que per mitjà de les marques (l'equivalent de les metadades dels fitxers binaris) s'hi incorpora informació relativa a com s'ha de representar el text, sobre quina estructura tenen les dades que conté, etc.
  - Combina dades i etiquetes que les marquen i que contenen informació addicional sobre l'estructura del text o la seva presentació.

# LLENGUATGE DE MARQUES

- Tipus de llenguatges de marques:
  - Llenguatges procedimentals (donar instruccions) i de presentació (definir format):
    - Orientats a especificar com s'ha de representar la informació.
    - Ex: HTML, CSS, etc..
  - Llenguatges descriptius o semàntics (Defineix estructura):
    - Orientats a descriure l'estructura de les dades que conté.
    - Ex: JSON, XML, etc...
  - Molts llenguatges combinen les característiques dels diferents tipus.

# LLENGUATGE DE MARQUES

- El llenguatge de marques és un dels sistemes per comunicar informació entre aplicacions. En aquesta informació també intervenen:
  - Dades
  - Emmagatzematge de dades

# LLENGUATGE DE MARQUES

- Les dades:
  - Són una representació del món real.
    - Per exemple:
      - les dimensions dels objectes que ens envolten
      - la nostra massa,
      - la velocitat del cotxe que acaba de passar
      - decibels del timbre de classe…
      - etc
  - Per si mateixa no constitueix informació perquè és el moment de processar-la que es genera.
  - Per processar-les cal emmagatzemar-les i un programa.
  - Un cop processades les dades, es poden prendre decisions.

# LLENGUATGE DE MARQUES

- Aspectes destacables de les dades:
  - El destinatari
  - Possibilitat de reutilitzar-les
  - La compartició d'aquestes

# LLENGUATGE DE MARQUES

- Destinatari:
  - Generalment els destinataris són:
    - Humans
      - En general tenen un estructura i un format. Per exemple: un títol, negreta, etc.
      - Processem les dades aplicant una intel·ligència que la immensa majoria dels casos, les màquines no tenen. (Excepció: IA)
    - Programes
      - Per poder tractar les dades de manera automàtica, és necessari que el programa pugui interpretar el format. Per exemple: el tipus de dada i el significat.

# LLENGUATGE DE MARQUES

- La reutilització de les dades:
  - Per reutilitzar les dades cal generar fitxers que poden ser de marques.
  - Els fitxers s'han de poder guardar a algun lloc.
  - L'error més habitual és generar fitxers per una tasca concreta. Fet que dificulta la reutilització.

# LLENGUATGE DE MARQUES

- La compartició de dades:
  - Per poder compartir les dades, cal guardar-les de manera que diferents sistemes les puguin reutilitzar.
    - Per exemple: Altres sistemes operatius, altres màquines, altres programes, etc

# LLENGUATGE DE MARQUES

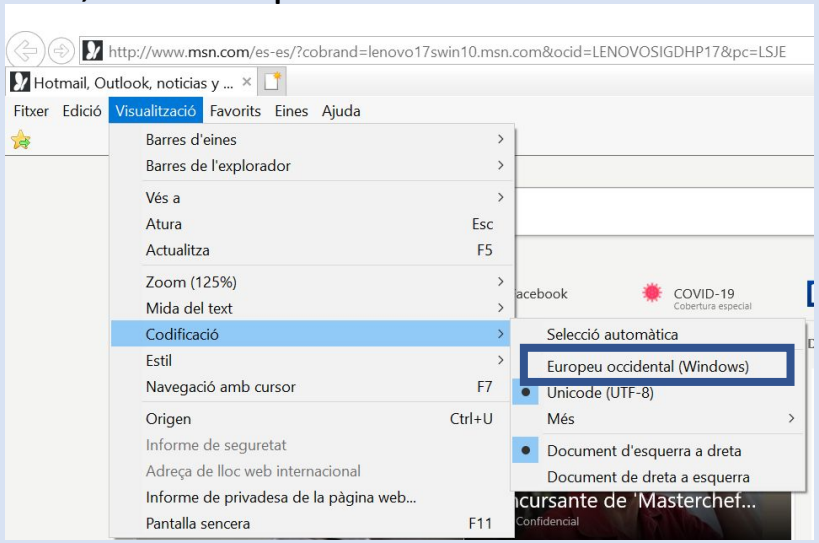
- Compartició d'informació:
  - Avantatges
    - Els sistemes de codificació permeten fàcilment la compartició d'informació.
    - Les dades poden llegir-les els éssers humans.
    - Les dades es poden reutilitzar per diversos programes com a editors o navegadors.
  - Desavantatges:
    - Ocupen més espai que les dades binàries.
    - Hi ha múltiples caràcters diferents.
    - No hi ha una estructura clara de dades per tal de tractar les dades.
    - Cada sistema operatiu fa un tractament diferent d'alguns caràcters com per exemple final de línia o salt de línia.

# LLENGUATGE DE MARQUES

- Emmagatzematge de dades:
  - Les dades es poden guardar en 2 formats:
    - Dades binàries:
      - Els avantatges són:
        - Es troben amb el mateix format que utilitza l'ordinador.
        - Estan optimitzades.
        - Els ordinadors les llegeixen més fàcilment.
        - Poden tenir estructura.
        - Es poden afegir metadades
          - Metadades
            - En aquest cas, informació descriptiva sobre el context, característiques, etc, de les pròpies dades.
            - La funció és facilitar la recuperació de les pròpies dades o avaluació, interoperatibitat, preservació o autentificació d'aquestes.

# LLENGUATGE DE MARQUES

- Emmagatzematge de dades:
  - Les dades es poden guardar en 2 formats:
    - Dades de text:
      - Moltes vegades, l'ésser humà és capaç d'interpretar-les.
      - Les dades s'escriuen caràcter a caràcter mitjançant un sistema de codificació.
      - Qualsevol sistema les pot interpretar si coneix el joc de caràcters en què està escrit.
      - Per defecte, en Linux el joc de caràcters és UTF-8, mentre que en Windows és ISO-8559.



# LLENGUATGE DE MARQUES

- Codis de caràcters:
  - El joc de caràcters que serveix de base a la resta és el codi ASCII:

| Caracteres ASCII de control | | |
|---|---|---|
| 00 | NULL | (carácter nulo) |
| 01 | SOH | (inicio encabezado) |
| 02 | STX | (inicio texto) |
| 03 | ETX | (fin de texto) |
| 04 | EOT | (fin transmisión) |
| 05 | ENQ | (consulta) |
| 06 | ACK | (reconocimiento) |
| 07 | BEL | (timbre) |
| 08 | BS | (retroceso) |
| 09 | HT | (tab horizontal) |
| 10 | LF | (nueva línea) |
| 11 | VT | (tab vertical) |
| 12 | FF | (nueva página) |
| 13 | CR | (retorno de carro) |

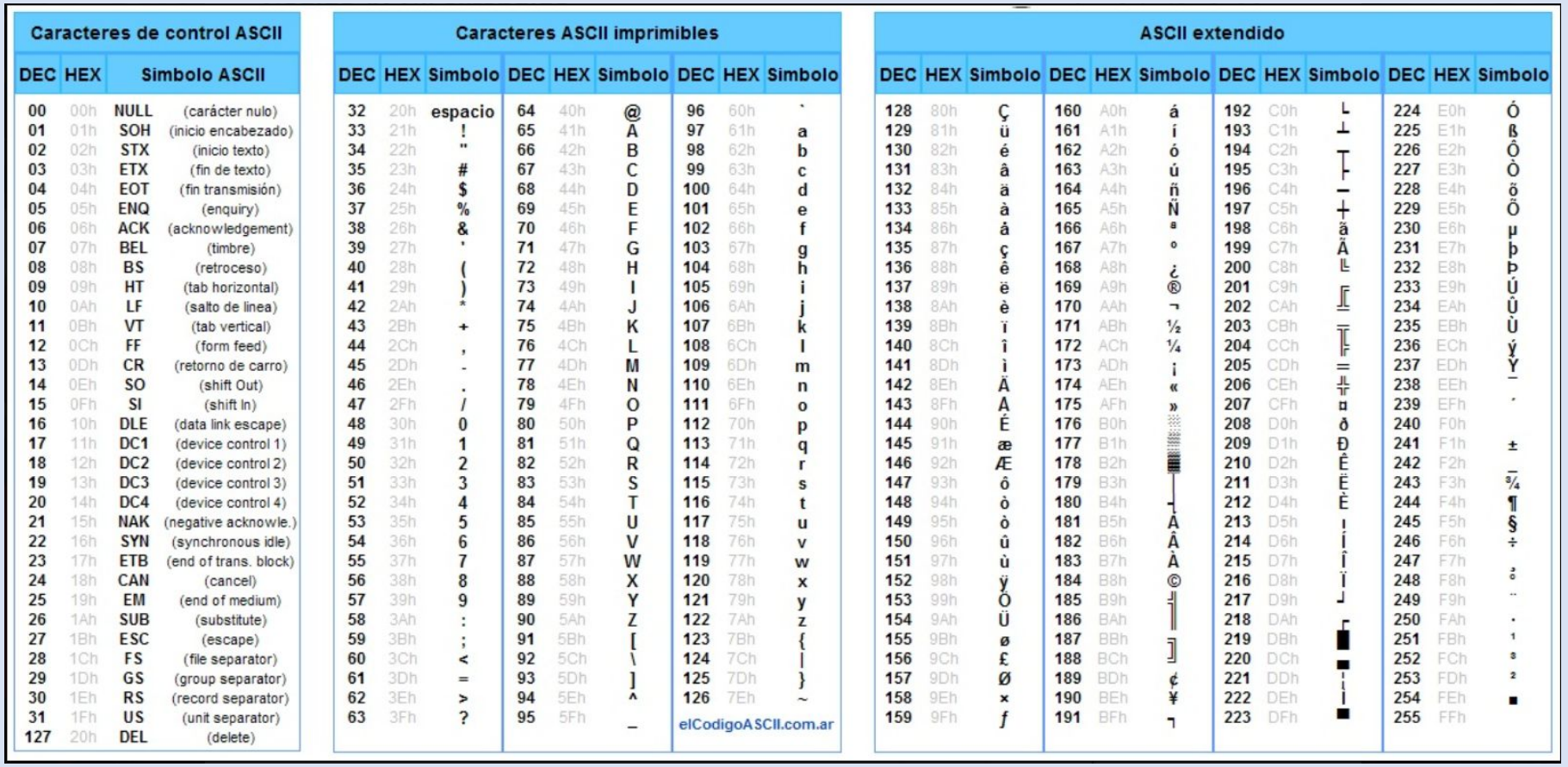
A un caràcter li correspon una codificació binaria.

Cada octet (o 7 bits) li correspon un caràcter o símbol de la llengua anglesa.
Problema: com a màxim 128 símbols  Es creen nous jocs com UTF-8 o ISO-8559 o ASCII estès (8 bits).

És necessari que la codificació sigui el més estàndard possible.

# LLENGUATGE DE MARQUES

- Codi ASCII estès:



**Caracteres de control ASCII**

| DEC | HEX | Simbolo ASCII | |
|---|---|---|---|
| 00 | 00h | NULL | (carácter nulo) |
| 01 | 01h | SOH | (inicio encabezado) |
| 02 | 02h | STX | (inicio texto) |
| 03 | 03h | ETX | (fin de texto) |
| 04 | 04h | EOT | (fin transmisión) |
| 05 | 05h | ENQ | (enquiry) |
| 06 | 06h | ACK | (acknowledgement) |
| 07 | 07h | BEL | (timbre) |
| 08 | 08h | BS | (retroceso) |
| 09 | 09h | HT | (tab horizontal) |
| 10 | 0Ah | LF | (salto de linea) |
| 11 | 0Bh | VT | (tab vertical) |
| 12 | 0Ch | FF | (form feed) |
| 13 | 0Dh | CR | (retorno de carro) |
| 14 | 0Eh | SO | (shift Out) |
| 15 | 0Fh | SI | (shift In) |
| 16 | 10h | DLE | (data link escape) |
| 17 | 11h | DC1 | (device control 1) |
| 18 | 12h | DC2 | (device control 2) |
| 19 | 13h | DC3 | (device control 3) |
| 20 | 14h | DC4 | (device control 4) |
| 21 | 15h | NAK | (negative acknowle.) |
| 22 | 16h | SYN | (synchronous idle) |
| 23 | 17h | ETB | (end of trans. block) |
| 24 | 18h | CAN | (cancel) |
| 25 | 19h | EM | (end of medium) |
| 26 | 1Ah | SUB | (substitute) |
| 27 | 1Bh | ESC | (escape) |
| 28 | 1Ch | FS | (file separator) |
| 29 | 1Dh | GS | (group separator) |
| 30 | 1Eh | RS | (record separator) |
| 31 | 1Fh | US | (unit separator) |
| 127 | 20h | DEL | (delete) |

**Caracteres ASCII imprimibles**

| DEC | HEX | Simbolo | DEC | HEX | Simbolo | DEC | HEX | Simbolo |
|---|---|---|---|---|---|---|---|---|
| 32 | 20h | espacio | 64 | 40h | @ | 96 | 60h | ` |
| 33 | 21h | ! | 65 | 41h | A | 97 | 61h | a |
| 34 | 22h | " | 66 | 42h | B | 98 | 62h | b |
| 35 | 23h | # | 67 | 43h | C | 99 | 63h | c |
| 36 | 24h | $ | 68 | 44h | D | 100 | 64h | d |
| 37 | 25h | % | 69 | 45h | E | 101 | 65h | e |
| 38 | 26h | & | 70 | 46h | F | 102 | 66h | f |
| 39 | 27h | ' | 71 | 47h | G | 103 | 67h | g |
| 40 | 28h | ( | 72 | 48h | H | 104 | 68h | h |
| 41 | 29h | ) | 73 | 49h | I | 105 | 69h | i |
| 42 | 2Ah | * | 74 | 4Ah | J | 106 | 6Ah | j |
| 43 | 2Bh | + | 75 | 4Bh | K | 107 | 6Bh | k |
| 44 | 2Ch | , | 76 | 4Ch | L | 108 | 6Ch | l |
| 45 | 2Dh | - | 77 | 4Dh | M | 109 | 6Dh | m |
| 46 | 2Eh | . | 78 | 4Eh | N | 110 | 6Eh | n |
| 47 | 2Fh | / | 79 | 4Fh | O | 111 | 6Fh | o |
| 48 | 30h | 0 | 80 | 50h | P | 112 | 70h | p |
| 49 | 31h | 1 | 81 | 51h | Q | 113 | 71h | q |
| 50 | 32h | 2 | 82 | 52h | R | 114 | 72h | r |
| 51 | 33h | 3 | 83 | 53h | S | 115 | 73h | s |
| 52 | 34h | 4 | 84 | 54h | T | 116 | 74h | t |
| 53 | 35h | 5 | 85 | 55h | U | 117 | 75h | u |
| 54 | 36h | 6 | 86 | 56h | V | 118 | 76h | v |
| 55 | 37h | 7 | 87 | 57h | W | 119 | 77h | w |
| 56 | 38h | 8 | 88 | 58h | X | 120 | 78h | x |
| 57 | 39h | 9 | 89 | 59h | Y | 121 | 79h | y |
| 58 | 3Ah | : | 90 | 5Ah | Z | 122 | 7Ah | z |
| 59 | 3Bh | ; | 91 | 5Bh | [ | 123 | 7Bh | { |
| 60 | 3Ch | < | 92 | 5Ch | \ | 124 | 7Ch | \| |
| 61 | 3Dh | = | 93 | 5Dh | ] | 125 | 7Dh | } |
| 62 | 3Eh | > | 94 | 5Eh | ^ | 126 | 7Eh | ~ |
| 63 | 3Fh | ? | 95 | 5Fh | _ | | | |

elCodigoASCII.com.ar

**ASCII extendido**

| DEC | HEX | Simbolo | DEC | HEX | Simbolo | DEC | HEX | Simbolo | DEC | HEX | Simbolo |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 128 | 80h | Ç | 160 | A0h | á | 192 | C0h | └ | 224 | E0h | Ó |
| 129 | 81h | ü | 161 | A1h | í | 193 | C1h | ┴ | 225 | E1h | ß |
| 130 | 82h | é | 162 | A2h | ó | 194 | C2h | ┬ | 226 | E2h | Ô |
| 131 | 83h | â | 163 | A3h | ú | 195 | C3h | ├ | 227 | E3h | Ò |
| 132 | 84h | ä | 164 | A4h | ñ | 196 | C4h | ─ | 228 | E4h | õ |
| 133 | 85h | à | 165 | A5h | Ñ | 197 | C5h | ┼ | 229 | E5h | Õ |
| 134 | 86h | å | 166 | A6h | ª | 198 | C6h | ã | 230 | E6h | µ |
| 135 | 87h | ç | 167 | A7h | º | 199 | C7h | Ã | 231 | E7h | þ |
| 136 | 88h | ê | 168 | A8h | ¿ | 200 | C8h | ╚ | 232 | E8h | Þ |
| 137 | 89h | ë | 169 | A9h | ® | 201 | C9h | ╔ | 233 | E9h | Ú |
| 138 | 8Ah | è | 170 | AAh | ¬ | 202 | CAh | ╩ | 234 | EAh | Û |
| 139 | 8Bh | ï | 171 | ABh | ½ | 203 | CBh | ╦ | 235 | EBh | Ù |
| 140 | 8Ch | î | 172 | ACh | ¼ | 204 | CCh | ╠ | 236 | ECh | ý |
| 141 | 8Dh | ì | 173 | ADh | ¡ | 205 | CDh | ═ | 237 | EDh | Ý |
| 142 | 8Eh | Ä | 174 | AEh | « | 206 | CEh | ╬ | 238 | EEh | ¯ |
| 143 | 8Fh | Å | 175 | AFh | » | 207 | CFh | ¤ | 239 | EFh | ´ |
| 144 | 90h | É | 176 | B0h | ░ | 208 | D0h | ð | 240 | F0h |  |
| 145 | 91h | æ | 177 | B1h | ▒ | 209 | D1h | Ð | 241 | F1h | ± |
| 146 | 92h | Æ | 178 | B2h | ▓ | 210 | D2h | Ê | 242 | F2h | ‗ |
| 147 | 93h | ô | 179 | B3h | │ | 211 | D3h | Ë | 243 | F3h | ¾ |
| 148 | 94h | ò | 180 | B4h | ┤ | 212 | D4h | È | 244 | F4h | ¶ |
| 149 | 95h | ò | 181 | B5h | Á | 213 | D5h | ı | 245 | F5h | § |
| 150 | 96h | û | 182 | B6h | Â | 214 | D6h | Í | 246 | F6h | ÷ |
| 151 | 97h | ù | 183 | B7h | À | 215 | D7h | Î | 247 | F7h | ¸ |
| 152 | 98h | ÿ | 184 | B8h | © | 216 | D8h | Ï | 248 | F8h | ° |
| 153 | 99h | Ö | 185 | B9h | ╣ | 217 | D9h | ┘ | 249 | F9h | ¨ |
| 154 | 9Ah | Ü | 186 | BAh | ║ | 218 | DAh | ┌ | 250 | FAh | · |
| 155 | 9Bh | ø | 187 | BBh | ╗ | 219 | DBh | █ | 251 | FBh | ¹ |
| 156 | 9Ch | £ | 188 | BCh | ╝ | 220 | DCh | ▄ | 252 | FCh | ³ |
| 157 | 9Dh | Ø | 189 | BDh | ¢ | 221 | DDh | ¦ | 253 | FDh | ² |
| 158 | 9Eh | × | 190 | BEh | ¥ | 222 | DEh | Ì | 254 | FEh | ■ |
| 159 | 9Fh | ƒ | 191 | BFh | ┐ | 223 | DFh | ▀ | 255 | FFh | |

# LLENGUATGE DE MARQUES

- Codi ASCII estès:
  - Hi ha diverses varietats d'ASCII especialitzades en una llengua en concret:
    - ISO 8859-1  Europa Oest
    - ISO 8859-2  Europa Central i Est
    - ISO 8859-3  Esperanto
    - …
    - ISO 8859-10  Àrea nòrdica
    - …
    - ISO 8859-15  Caràcters especials

# LLENGUATGE DE MARQUES

- Codi ASCII estès:
  - Inconvenients:
    - Únicament es va generar per idiomes que utilitzen l'alfabet llatí.
    - Cada idioma afegeix els seus caràcters particulars a l'ASCII expandit.

# LLENGUATGE DE MARQUES

- Unicode:
  - És una alternativa al codi ASCII estès.
  - Serveix per a totes les llengües.
  - Permet afegir codis llatins.
  - Per cada símbol hi ha un identificador únic.
  - 3 formes de definició bàsica:
    - UTF-8 (UTF  Unicode Transformation Format)
    - UTF-16
    - UTF-32
  - Unicode utilitza caràcters de 8, 16 o 32 bits en funció de la representació específica  documents Unicode solen requerir fins al doble d'espai en disc que els documents ASCII o Latin-1 (ISO-8559-1).
  - Els primers 256 són iguals que Latin-1.

# LLENGUATGE DE MARQUES

- Per a que els programes puguin tractar les dades  s'han d'estructurar.
  - Exemples per estructurar dades:
    - Generar a un fitxer les dades en format CSV.
    - Generar fitxers de marques.

# LLENGUATGE DE MARQUES

- Fitxers de Marques:
  - Són fitxers emmagatzemats amb algun codi de caràcters estàndard (UTF-8, ASCII, etc).
  - Es podem tractar amb editors.
  - Es poden afegir metadades.
  - Defineixen l'estructura de les dades.

# LLENGUATGE DE MARQUES

- Marques:
  - Són un conjunt 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.
  - Són anotacions que s'incorporen a les dades però que no en formen part.
  - A més, han de ser distingibles del text normal.
  - Exemple:

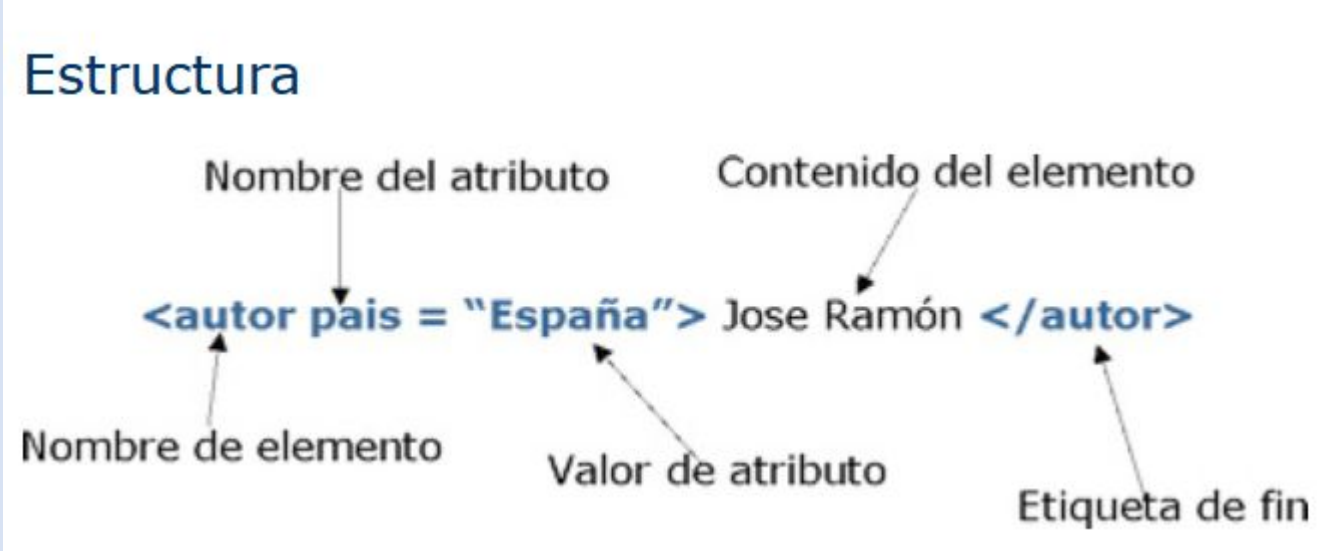
```
<?xml version="1.0" encoding="UTF-8"?>
<response>
 <row>
  <row _id="row-78q9_hsqr.nj23" _uuid="00000000-0000-0000-4705-26D44C26AE32" _position="0" _address=
 "https://analisi.transparenciacatalunya.cat/resource/zzzz-zzzz/row-78q9_hsqr.nj23">
   <n_m_registre>77</n_m_registre>
   <nom_entitat>APA CN MIXTO LES TERMES</nom_entitat>
   <adre_a>Illa Bella, s/n</adre_a>
   <cp>08202</cp>
   <municipi>Sabadell</municipi>
   <comarca>Vallès Occidental</comarca>
   <tel_fon_faxs>
    <tel_fon_fax>93-7106594</tel_fon_fax>
   </tel_fon_faxs>
   <tipus_entitat>AMPA</tipus_entitat>
   <modalitats>
    <modalitat>Basquetbol</modalitat>
   </modalitats>
   </row>
```

# LLENGUATGE DE MARQUES

- Característiques del llenguatge de marques:
  - Que es basen en el text pla.
    - Els caràcters del text poden estar codificats en diferents codis de caràcters: ASCII, ISO-8859-1, UTF-8, etc.
  - Que permeten fer servir metadades.
    - Donen informació de les dades contingudes.
  - Que són fàcils d'interpretar i processar.
    - Permeten automatitzar processos pel tractament de les dades del fitxer de marques.
  - Que són fàcils de crear i prou flexibles per representar dades molt diverses.
    - Permeten crear text, imatges, pàgines web, fórmules matemàtiques, etc.

# XML

- XML (eXtensible Markup Language/llenguatge d'etiqueta extensible):
  - És un llenguatge estàndard, una recomanació del World Wide Web Consortium (W3C, www.w3.org/TR/REC-xml)
  - Les etiquetes són l'element essencial. El funcionament de les etiquetes és el següent:



# XML

- XML:
  - Permet separar el contingut de la manera de com serà representat.
  - Sempre una etiqueta s'ha de tancar.
  - Hi ha 2 tipus d'etiquetes:
    - Les etiquetes d'obertura
    - Les etiquetes de tancament
  - Es recomana triar els noms de les etiquetes de manera que puguin ser interpretades per qui les llegirà independentment de si les processa una persona o programa.

# XML

- XML:
  - Alguns caràcters donen problemes a les etiquetes però es poden substituir:

| Símbol | Substitució |
|---|---|
| < | &lt; |
| > | &gt; |
| " | &quot; |
| ' | &apos; |
| & | &amp; |

# XML

- XML:
  - CDATA:
    - Serveix per afegir un element HTML o codi, un element de marques no ben formatejat o grans fragments de codi.

```
<![CDATA[
   characters with markup
]]>
```

# XML

- XML:
  - Atributs:
    - És la manera d'afegir contingut al fitxer XML.
    - Només s'especifiquen a l'etiqueta d'obertura.
    - Tots els valors dels atributs han d'estar envoltats de cometes.

# XML

- XML:
  - Regles per definir un nom vàlid a les etiquetes:
    - Els noms han de començar per una lletra de l'alfabet, el caràcter de subratllat (_) o un guió (-). També s'accepta el caràcter de dos punts (:), però està reservat.
    - Els caràcters en majúscules són diferents dels caràcters en minúscules.
    - No hi pot haver espais enmig del nom.
    - No poden començar per la paraula XML tant si qualsevol de les lletres està en majúscules o en minúscules. Aquestes paraules es reserven per a estandarditzacions futures.

# XML

- XML:
  - Exemples de posar noms:

| Etiquetes correctes | Etiquetes incorrectes |
|---|---|
| <correu1/> | <1Correu/> |
| <correu-electronic/> | <correu electronic/> |
| <element /> | < element/> |
| <_Carai/> | <%descompte/> |
| <svg:rectangle /> | <xml-rectangle/> |

# XML

- XML:
  - Verificació del format d'un document XML:
    - Comprovar si un document XML està ben formatejat, és comprovar que no s'incompleixen cap de les regles de definició d'XML.
    - Les regles són:
      - Només hi pot haver un element arrel.
        - Element arrel  És l'etiqueta que surt al primer lloc.
      - Totes les etiquetes que s'obren s'han de tancar.
      - Les etiquetes han d'estar imbricades correctament. (Exemple següent pestaña)
        - No es pot tancar una etiqueta si encara hi ha una etiqueta que forma part del contingut que no ha estat tancada. Per facilitar aquesta regla es recomana sagnar el document.
      - Els noms de les etiquetes han de ser correctes.
      - Els valors dels atributs han d'estar entre cometes.

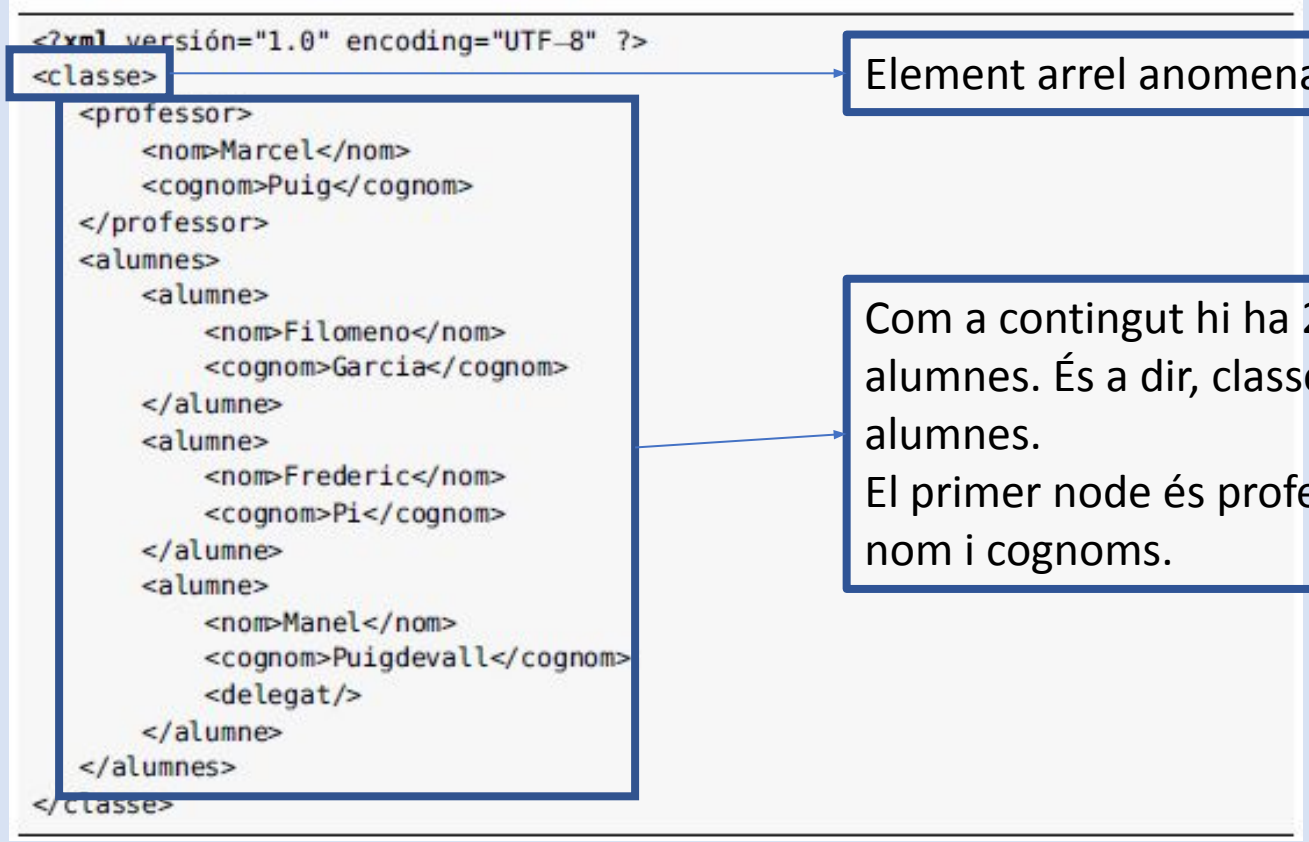
# XML

- XML:
  - Exemple de document sagnat:

```
<institut>
    <classe>
        <alumne>Pere</alumne>
        <alumne>Joan</alumne>
    </institut>
</classe>
```

# XML

- XML:
  - La forma d'estructurar el document és mitjançant una jerarquia format per nodes.



# XML

- XML:
  - Editor: