1.- Esquemes i vocabularis en XML

Suposem el següent problema:

- Imagina que estem gestionant una base de dades d'una biblioteca en format XML.
- La informació que volem registrar per a cada llibre són: **autor**, **títol** i **any** de publicació.
- Els llibres els cataloguen varies persones, i, com no s'han posat d'acord, cada catalogador realitza diferents versions.

- **Versió 1:**

```
<llibre>
    <títol>El Arte de la Guerra</títol>
    <autor>Sun Tzu</autor>
    <publicació>200 a.C.</publicació>
</llibre>
```

- **Versio 2:**

```
<libro>
    <título>El Arte de la Guerra</título>
    <publicación>200 a.C.</publicación>
    <autor>Sun Tzu</autor>
</libro>
```

- **Versió 3:**

```
<obra>
    <nombre>El Arte de la Guerra</nombre>
    <escritor>Sun Tzu</escritor>
    <lanzamiento>200 antes de Cristo</lanzamiento>
</obra>
```

Al poc de temps, algú decideix agregar més elements:

- **Versió 1 ampliada:**

```
<llibre>
    <títol>El Arte de la Guerra</títol>
    <autor>Sun Tzu</autor>
    <publicació>200 a.C.</publicació>
    <editorial>Ediciones Z </editorial >
</llibre>
```

Com veiem als exemples, cada persona ha creat la seua pròpia estructura i nomenclatura, el que dificulta la comprensió i la interpretació de les dades dels llibres, es a dir **falta Coherència**.

Esta clar que ninguna de les anteriors versions en el exemple es compatible amb les altres, i per tant seria crític **posar-se d'acord amb el que es pot fer, el que pot aparèixer i en quin ordre hem de fer-ho.**

XML és com un idioma que les computadores i els sistemes d'informació utilitzen per comunicar-se. No obstant això, perquè aquesta comunicació sigui eficaç, tots han de parlar el mateix idioma.
Aquí és on intervenen els "**esquemes**" i "**vocabularis**" **XML**.

- Els **"esquemes" XML** són com formularis que defineixen **l'estructura** i **els tipus de dades** que s'han d'utilitzar en els documents XML. Estableixen regles perquè les dades siguin coherents i comprensibles.

- **Els "vocabularis" XML** són com el conjunt de paraules i frases que hem d'utilitzar. Defineixen les **etiquetes** i **elements específics** que podem usar en els teus documents XML.

Per establir l'estructura d'un document XML s'utilitzen **"llenguatges de definició de vocabularis"** o **"llenguatges d'esquemes"**, els més coneguts són **DTD, XML Schema i Relax NG**.

En resum ...

- Un document XML pot:
  - **Estar ben format:** si compleix amb la **sintaxi** XML.
  - **Ser vàlid:** Si a més d'estar ben format **compleix** determinades **regles** i **normes**.
- Per a **establir les regles** de construcció en XML utilitzem **DTD i XML Schema**
- **Validar** és un procés habitual, sobre tot quan es comparteix informació entre sistemes. Esta validació la realitzem amb programes especials "**processadors**" o "**validadors**" també conegut com a **"parsers"**)

2.- DTD

DTD: Document Type Definitions – Validació i Definició de Documents

Un **DTD** és una especificació que defineix **l'estructura**, els **elements** i els **atributs** permesos en documents XML, proporcionant una guia per validar i interpretar correctament aquests documents

El DTD ens permet crear el nostre propi llenguatge de marcat per a aplicacions específiques. Defineix:

- Tipus d'elements
- Atributs
- Entitats permeses
- També es poden expressar restriccions.

Podem crear DTD's de dos maneres:

- **Fitxer extern**: Pot ser compartit per diversos (milers?) de documents
- **En el propi document XML**: Com a part de la seva declaració de tipus de document. (S'anomena **inline**)

• Sintaxi básica:

```
<!DOCTYPE element DTD identifier
[
   declaration1
   declaration2
   ........
]>
```

- **<!DOCTYPE:** indica que estem declarant un tipus de document DTD.
- **element**: especifica el nom de l'element que serà la arrel o element principal del document XML.
- **DTD identifier**: Aquest és un identificador que fa referència a la definició del tipus de document. Pot ser una ruta d'accés a un fitxer local o una URL que apunta a un fitxer DTD extern. Quan es fa referència a un fitxer extern, es parla de "Subconjunt Extern."
- **[ ]:** Els parèntesis quadrats indiquen que es pot incloure una llista opcional de declaracions d'entitats, que formen part del subconjunt intern del DTD.

Per exemple:

- **DTD Interna:** El DTD es defineix dins del propi document XML. S'afegeix una secció DTD entre **<!DOCTYPE** i **>** just després de la declaració **XML**

```
<?xml version="1.0" encoding="UTF-8" standalone="yes" ?>

<!DOCTYPE articles [
        <!ELEMENT articles (article+)>
        <!ELEMENT article (titol, autor, contingut)>
        <!ELEMENT titol (#PCDATA)>
        <!ELEMENT autor (#PCDATA)>
        <!ELEMENT contingut (#PCDATA)>
        ]>

<articles>
    <article>
        <titol>Article 1</titol>
        <autor>Inés Tornudo </autor>
        <contingut>Este és el contingut de l'Article 1.</contingut>
    </article>
    <!-- Altres articles aquí -->
</articles>
```

- **DTD Externa**: En aquest cas, el DTD es defineix en un fitxer separat i s'associa amb el document XML mitjançant una **referència** a la secció DTD.

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>

<!DOCTYPE articles SYSTEM "articles.dtd">

<articles>
    <article>
        <titol>Article 1</titol>
        <autor>Inés Tornudo </autor>
        <contingut>Este és el contingut de l'Article 1.</contingut>
    </article>
    <!-- Altres articles aquí -->
</articles>
```

articles.dtd

```
<!ELEMENT articles (article+)>
<!ELEMENT article (titol, autor, contingut)>
<!ELEMENT titol (#PCDATA)>
<!ELEMENT autor (#PCDATA)>
 <!ELEMENT contingut (#PCDATA)>
```

El document també pot estar en una ubicació externa:

```
<!DOCTYPE articles PUBLIC "https://exemple.com/dtd/articles.dtd">
```

En els dos casos les regles estan en un fitxer amb extensió **.dtd**. Farem referència al seu **nom** i **direcció** (URI). Pot ser:

- **Privada**. És la més comuna. La definim nosaltres. S'utilitza amb **SYSTEM**.
- **Pública**. La defineix un organisme d'estandardització. S'utilitza amb **PUBLIC**.

• Estructura DTD

En un DTD es poden declarar:

- ✓ **Elements**
- ✓ **Atributs**
- ✓ **Entitats**
- ✓ **Notacions**

Recorda: Un document XML serà **vàlid** si, a més de **no** tindre **errors** de sintaxis, **compleix** l'indicat en les **declaracions d'elements**, **atributs**, **entitats** i notacions del **DTD** associat a eixe XML.

• Declaració d'Elements

Utilitzarem la següent sintaxis:

```
<!ELEMENT nom-element tipus-de-contingut>
```

On:

- **nom_element**: És el nom de l'element que estem definint.
- **tipus_contingut**: Indica quin tipus de contingut pot tenir l'element (com ara "PCDATA" , per a text, "EMPTY" per a elements buits, "ANY" per a qualsevol tipus de contingut o un altre element o seqüència d'elements).

Per exemple:

```
<!ELEMENT receta (titulo, ingredientes, procedimiento)>
```

Seguint la definició de l'element anterior...

XML Vàlid	XML NO Vàlid

`<receta>`
`  <titulo>...</titulo>`
`  <ingredientes>...</ingredientes>`
`  <procedimiento>...</procedimiento>`
`</receta>`
	`<receta>`
	`  <parrafo>Això és un paràgraf</parrafo>`
	`  <titulo>...</titulo>`
	`  <ingredientes>...</ingredientes>`
	`  <procedimiento>...</procedimiento>`
	`</receta>`

• Tipus de Continguts

Els Tipus de Continguts dels Elements poden ser:

- Text: **(#PCDATA)**.

```
<!ELEMENT cotxe (#PCDATA)>
```

- **EMPTY** (Buit). Pot no tindre contingut, sol usar-se per als atributs

```
<!ELEMENT linia_de_separacio EMPTY>
```

- **ANY** (text i altres elements): . Pot tindre qualsevol contingut (No es sol usar**)**

```
<!ELEMENT batiburrillo ANY>
```

- **Mixted** (text i altres elements): Pot tenir caràcters de tipus dades o una mesclade caràcters i sub-elements especificats

```
<!ELEMENT enfasis (#PCDATA)>
<!ELEMENT parrafo (#PCDATA|enfasis)*>
```

- **Tipus Element:** Sols pot contenir sub-elements que consten a l'especificació de contingut.

```
<!ELEMENT article (titol, autor, contingut)>
```

```
<!DOCTYPE articles [
        <!ELEMENT articles (article+)>
        <!ELEMENT article (titol, autor, contingut)>
        <!ELEMENT titol (#PCDATA)>
        <!ELEMENT autor (#PCDATA)>
        <!ELEMENT contingut (#PCDATA)>
        ]>
```

Per a que un tipus d'element tinga contingut d'elements s'especifica un **model de contingut**.

• Models de Continguts

- **Identificador General**: Indica que <aviso> només pot contenir un sol **<parrafo>**

```
<!ELEMENT aviso (parrafo)>
```

- **Seqüència**: La coma **","** denota una seqüència. **<aviso>** ha de contenir un **<titulo>** seguit d'un **<parrafo>.**

```
<!ELEMENT aviso (titulo, parrafo)>
```

- **Opció** : La barra vertical **"|"** indica una opció. **<aviso>** pot contenir o bé un **<parrafo>** o bé un **<grafico>**.

```
<!ELEMENT aviso (parrafo | grafico)>
```

  - El nombre d'opcions no està limitat a dues, i es poden agrupar usant parèntesis.

```
<!ELEMENT aviso (titulo, (parrafo | grafico))>
```

En aquest últim cas, el **<aviso>** ha de contindre un **<titulo>** seguit d'un **<parrafo>** o d'un **<grafico>.**

• Quantificadors

També odem utilitzar quantificadors de freqüència, com **"*"** (zero o més vegades), **"+"** (una o més vegades) i **"?"** (zero o una vegada) per especificar la quantitat d'instàncies d'un element o subelement.

En resum:

(?)	= 0, 1 elemento
(*)	= 0 ó más elementos
(+)	= 1 ó más elementos
(|)	= alternativa
(,)	= secuencia
EMPTY	= vacío
ANY	= cualquier estructura de subelementos
#PCDATA	= cadena de caracteres analizados

Exemple Models de Continguts

```
<!ELEMENT aviso (titulo?, (parrafo+, grafico))>
```

- **<aviso>:** Aquest és **l'element principal** que estem definint.

La seua estructura és:

- **(titulo?, (parrafo+, grafico))**: Defineix l'estructura interna de l'element <aviso>.
    - **(titulo?, ...):**
        - **<titulo>:** Este element és opcional (indicat per ?), es a dir, pot aparèixer **zero** o **una volta com a molt** dins de l'element <aviso>. No és obligatori tenir un **<titulo>** en cada **<aviso>.**
    - **(parrafo+, grafico):** Esta part defineix que dins de l'element **<aviso>,** hi ha una seqüència de paràgrafs **<parrafo>** seguida d'un únic element **<grafico>.**

      Les regles són les següents:
        - **<parrafo>+:** Esta part indica que hi ha un o més elements **<parrafo>** (**mínim un**) dins de l'element **<aviso>.** Es a dir, hi ha almenys un paràgraf **<parrafo>.**
        - **<grafico>:** Aquest element **<grafico>** ha de estar present després de la seqüència de paràgrafs **<parrafo>.** És a dir, hi haurà un únic element **<grafico>** després dels paràgrafs.

Això defineix una estructura flexible per a l'element <aviso>, on pots tenir zero o un **<titulo>,** seguit de paràgrafs **(<parrafo>)** i, finalment, un element **<grafico>.** També pots tenir múltiples paràgrafs abans del <**grafico>.** Aquesta declaració permet diversos formats de <aviso>, com ara:

- **<aviso>**
- **<aviso><titulo></titulo><parrafo></parrafo><parrafo></parrafo><grafico></grafico>**
- **<aviso><parrafo></parrafo><grafico></grafico>**

I així successivament.