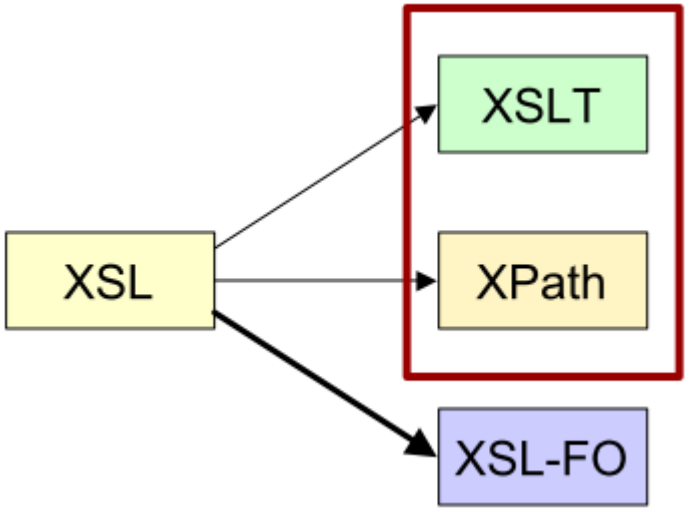
Unidad 5: Conversión y adaptación de documentos XML

XPATH

1. INTRODUCCIÓN

Los documentos XML son documentos de texto con etiquetas que contienen exclusivamente información sin entrar en detalles de formato. Esto implica la necesidad de algún medio para expresar la transformación de un documento XML para que una persona pueda utilizar directamente los datos para leer, imprimir, etc

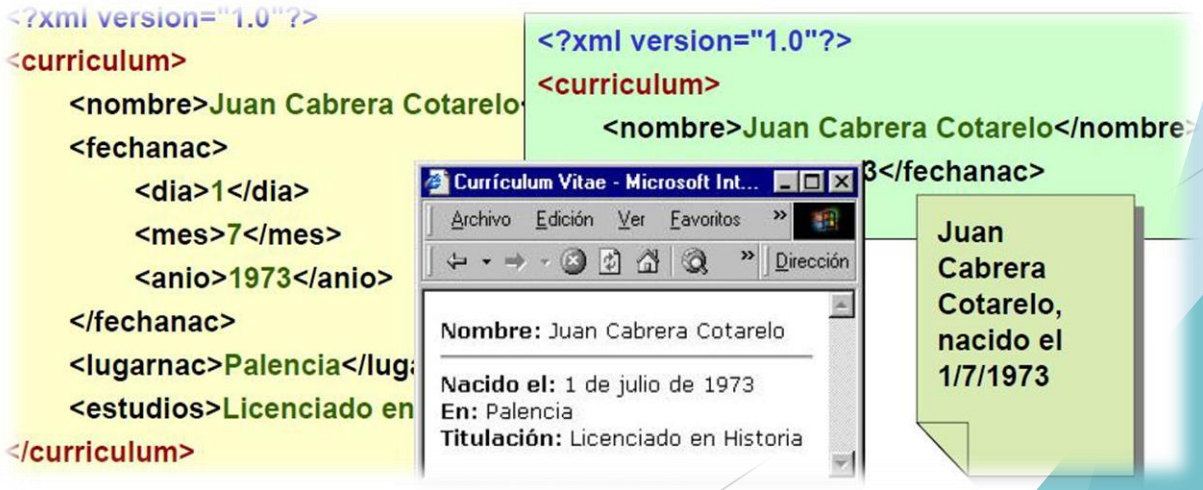
XSL(eXtensible Stylesheet Language) es un especificación de W3C consistente en un estándar de hojas de estilo especialmente creado para la notación XML. En realidad es una combinación de los siguientes estándares:



1. INTRODUCCIÓN

- **XSLT:** permite definir el modo de transformar un documento XML en otro.
- **XSL-FO:** Se utiliza para transformar XML en un formato imprimible y legible por una persona, por ejemplo en un documento PDF.
- **Xpath:** permite el acceso a los diversos componentes de un documento XML

Una transformación proporciona un conjunto de reglas para convertir datos descritos en un conjunto de etiquetas, en datos descritos en otro conjunto de etiquetas.



XPATH

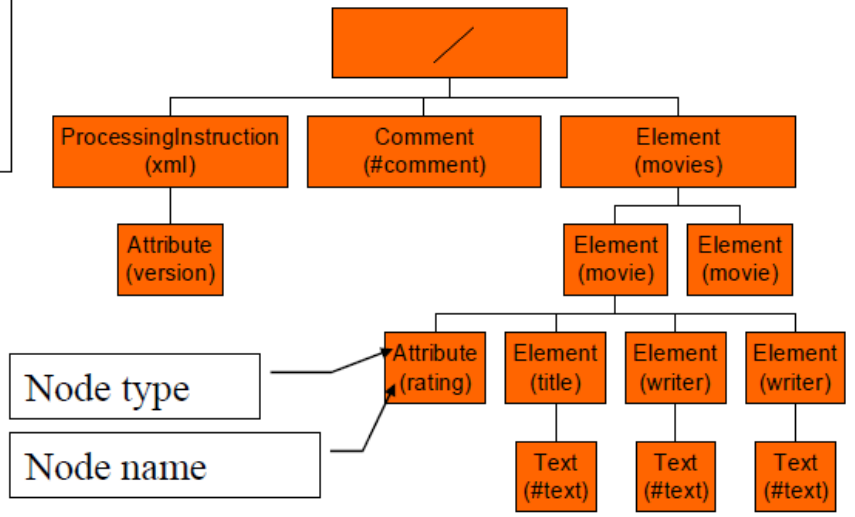
XML Path Languaje

2. XPATH

Es un estándar diferente de XML, aprobado por el W3C, que nos permite acceder a **partes de un documento XML basándonos en las relaciones** de parentesco entre los nodos del documento.

Su estilo de notación es similar a las rutas de los ficheros, pero se refiere a nodos en un documento XML. Ejemplo: /fecha/día

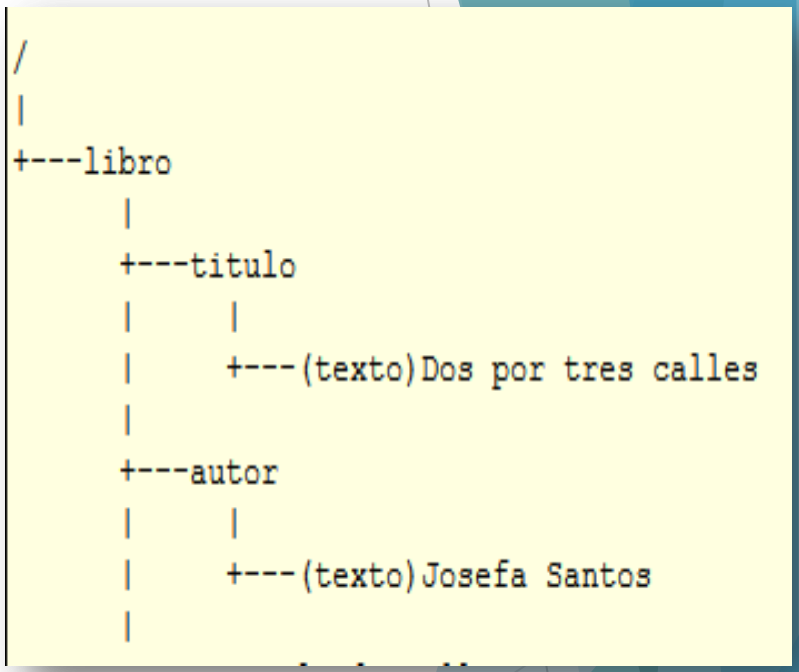
```
<?xml version="1.0" ?>
<!-- Mis peliculas preferidas -->
<movies>
  <movie rating="PG-13" >
    <title>Raising Arizona</title>
    <writer>Ethan Coen</writer>
    <writer>Joel Coen</writer>
</movie>
<movie> .... </movie>
</movies>
```



2.1 TERMINOLOGÍA, TIPOS DE NODOS

NODO RAÍZ

Se identifica por **/**. No se debe confundir el nodo raíz con el elemento raíz del documento. Así, si el documento XML de nuestro ejemplo tiene por elemento raíz a libro, éste será el primer nodo que cuelgue del nodo raíz del árbol, el cual es: **/**



/ hace referencia al nodo raíz del árbol, pero no al elemento raíz del documento XML, por más que un documento XML solo pueda tener un elemento raíz. De hecho, podemos afirmar que el nodo raíz del árbol contiene al elemento raíz del documento.

2.1 TERMINOLOGÍA, TIPOS DE NODOS

NODOS ELEMENTO

Son cada uno de los elementos del documento XML. Todos ellos tienen un elemento padre. El padre del elemento raíz, es el nodo raíz del documento.

Los nodos elemento tienen a su vez hijos, que son:

- nodos elemento
- nodos texto
- nodos comentario
- nodos de instrucciones de proceso.

Los nodos elemento también tienen propiedades tales como su nombre, sus atributos e información sobre los "espacios de nombre" que tiene activos.

2.1 TERMINOLOGÍA, TIPOS DE NODOS

NODOS TEXTO

Son aquellos caracteres del documento que no están marcados con ninguna etiqueta. No tienen hijos.

NODOS ATRIBUTO

No se consideran hijos del elemento al que están asociados sino etiquetas añadidas al nodo elemento.

NODOS DE COMENTARIO Y DE INSTRUCCIONES DE PROCESO

Son los nodos que se generan para los elementos con comentarios e instrucciones de proceso.

NODO ACTUAL

Es aquél al que nos referimos cuando se evalúa una expresión en XPath.

2.2 RESULTADOS Y TOKENS

Una expresión XPath permite extraer uno o varios fragmentos de información de un documento XML.

Los resultados posibles son:

- Conjunto o lista de nodos
- Valor booleano
- Número
- Texto

Ejemplo consulta:

```
/libro/capitulo/parrafo
```

- Hace referencia a TODOS los elementos parrafo que cuelguen directamente de CUALQUIER elemento capitulo que cuelgue de CUALQUIER elemento libro que, finalmente, cuelguen del nodo raíz, /.
- Una expresión XPath nos devuelve una lista de apuntadores a los elementos que encajan en el patrón. Dicha lista puede estar vacía o contener uno o más nodos.

2.2 RESULTADOS Y TOKENS

Los **tokens** que podemos usar en XPATH son:

- Paréntesis, “()”; llaves , “{}” y corchetes, “[ ]”.
- Elemento actual, elemento padre.
- Atributo, “@”.
- Elemento, “*”.
- Separador, “::”.
- Coma, “,”.
- El nombre de un elemento.
- Operadores: and, or, mod, div, *, /, //, |, +, -, =, !=, <, >, <=, >=.
- Nombres de función.
- Denominación de ejes: ancestor, ancestor-or-self-atribute, child, descendant, descendant-or-self, following, following-sibling, namespace, parent, preceding, preceding-sibling, self.
- Literales, se ponen entre comillas dobles o simples. Pueden anidarse alternando el tipo de comillas.
- Números.
- Referencias a variables, para lo que se utiliza la sintaxis: $nombreVariable

Ejemplo XPATH

```
<?xml version="1.0"?>
<libro>
  <titulo>Dos por tres calles</titulo>
  <autor>Josefa Santos</autor>
  <capitulo num="1">
    La primera calle
    <parrafo>Era una sombria noche del mes de agosto...</parrafo>
    <parrafo destacar="si">Ella, inocente cual
      <enlace href="http://www.enlace.es">mariposa</enlace>
      que surca el cielo en busca de libaciones...</parrafo>
  </capitulo>
  <capitulo num="2" public="si">
    La segunda calle
    <parrafo>Era una obscura noche del mes de septiembre...</parrafo>
    <parrafo>Ella, inocente cual
      <enlace href="http://www.abejilla.es">abejilla</enlace>
      que surca el viento en busca del nectar de las flores...
    </parrafo>
  </capitulo>
  <apendice num="a" public="si">
    La tercera calle
    <parrafo>Era una densa noche del mes de diciembre...</parrafo>
    <parrafo>Ella, candida cual
      <enlace href="http://www.pajarillo.es">elefante</enlace>
      que surca el espacio en busca de bichejos para
    comer...</parrafo>
  </apendice>
</libro>
```

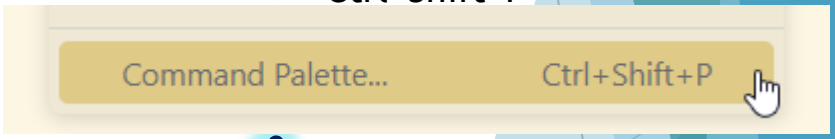
Evaluar XPath

VS Code

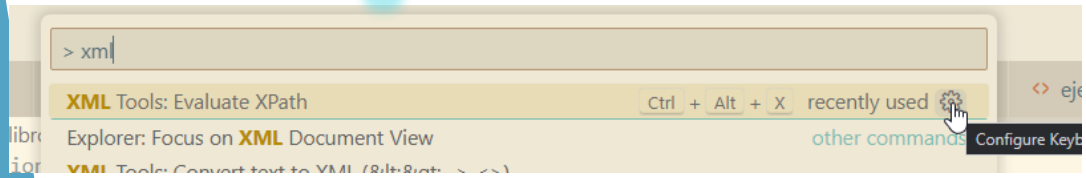
Instalar

plugin

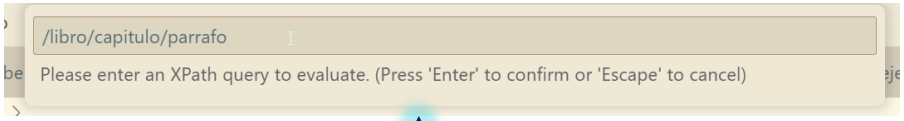
Click derecho sobre un XML y elegir en el menú contextual
o
Ctrl+Shift+P



Usa el buscador para acceder a "XML Tools: Evaluate XPath"
Crea un atajo de teclado para próximas ejecuciones



Introduce la consulta XPath en el cuadro de texto y pulsa enter.
Después de cada consulta, cierra la consola de resultados para un correcto funcionamiento



2.3 EJES

Los **ejes**, permiten seleccionar el subárbol dentro del nodo contexto que cumple un patrón. Pueden ser o no de contenido.

- **child**, es el eje utilizado por defecto. Su forma habitual es la barra, “/”, aunque también puede ponerse: /child::
- **attribute**, permite seleccionar los atributos que deseemos. Es el único eje que veremos que no es de contenido.
- **descendant**, permite seleccionar todos los nodos que descienden del conjunto de nodos contextos. Se corresponde con la doble barra, //, aunque se puede usar: descendant::
- **self**, se refiere al nodo contexto y se corresponde con el punto ”.”
- **parent**, selecciona los nodos padre, para referirnos a él usamos los dos puntos, “..”.
- **ancestor**, devuelve todos los nodos de los que el nodo contexto es descendiente.

Ejemplo consulta: /child::libro/child::autor/child::nombre/attribute::apellido

Ejemplo visual

2.4 SINTAXIS ABREVIADA

Una expresión XPath es una cadena de texto que representa un recorrido en el árbol del documento. Evaluar una expresión XPath es buscar si hay nodos en el documento que se ajustan al recorrido definido en la expresión. El resultado de la evaluación son todos los nodos que se ajustan a la expresión.

Las expresiones XPath se pueden escribir de dos formas distintas:

- **sintaxis abreviada**: más compacta y fácil de leer (utilizaremos esta)
- **sintaxis completa:** más larga pero con más opciones disponibles

Completa	Abreviada
child::	No es necesario
attribute::	@
descendant-or-self::node()/	//
self::node()	.
parent::node()	..

/child::libro/child::autor/child::nombre/attribut::apellido
equivale a
/libro/autor/nombre/@apellido

Usando eje **attribute::**, que puede sustituirse por la arroba, “@”.

2.5 ATRIBUTOS

```
<?xml version="1.0" encoding="iso-8859-1"?>
<agenda>
    <propietario>
        <identificadores>
            <nombre>Alma</nombre>
            <apellidos>López Terán</apellidos>
        </identificadores>
        <direccion>
            <calle>El Percebe 13, 6F</calle>
            <localidad>Torrelavega</localidad>
            <cp>39300</cp>
        </direccion>
        <telefonos>
            <movil>970898765</movil>
            <casa>942124567</casa>
            <trabajo>628983456</trabajo>
        </telefonos>
    </propietario>
    <contactos>
        <persona id="p01">
            <identificadores>
                <nombre>Inés</nombre>
                <apellidos>López Pérez</apellidos>
            </identificadores>
            <direccion>
                <calle>El Ranchito 24, 6B</calle>
                <localidad>Santander</localidad>
                <cp>39006</cp>
            </direccion>
            <telefonos>
                <movil>970123123</movil>
            </telefonos>
        </persona>
        . . .
        <persona id="p02">
            <identificadores>
                <nombre>Roberto</nombre>
                <apellidos>Gutiérrez Gómez</apellidos>
            </identificadores>
            <direccion>
                <calle>El Marranito 4, 2F</calle>
                <localidad>Santander</localidad>
                <cp>39004</cp>
            </direccion>
            <telefonos>
                <movil>970987456</movil>
                <casa>942333323</casa>
            </telefonos>
        </persona>
        <persona id="p03">
            <identificadores>
                <nombre>Juan</nombre>
                <apellidos>Sánchez Martínez</apellidos>
            </identificadores>
            <direccion>
                <calle>El Cangrejo 10, sn</calle>
                <localidad>Torrelavega</localidad>
                <cp>39300</cp>
            </direccion>
            <telefonos>
                <movil>997564343</movil>
                <casa>942987974</casa>
                <trabajo>677899234</trabajo>
            </telefonos>
        </persona>
    </contactos>
</agenda>
```

2.6 ATRIBUTOS

Nombre del propietario de la agenda.

/agenda/propietario/identificadores/nombre

Teléfono de casa del propietario.

/agenda/propietario/telefonos/casa

Nombres y apellidos de los contactos de la agenda.

//contactos/persona/identificadores/nombre | //contactos/persona/identificadores/apellidos

Nombre e identificador de cada contacto.

//contactos/persona/identificadores/nombre | //contactos/persona/@id

Datos del contacto con identificador "p02".

//contactos/persona[@id="p02"]/*/*

Identificadores de los contactos que tienen teléfono en casa.

//contactos/persona/telefonos/casa/../../@id

2.7 NODOS

Nodos test, permiten restringir lo que devuelve una expresión XPath.
Podemos agruparlos en función de los ejes a los que se puede aplicar.

Aplicable a cualquier eje:

- **“*”**, solo devuelve elementos, atributos o espacios de nombres pero no permite obtener nodos de texto, o comentarios de cualquier tipo.
- **node(),** devuelve los nodos de todos los tipos.

Solo aplicables a ejes de contenido:

- **text(),** devuelve cualquier nodo de tipo texto.
- **comment(),** devuelve cualquier nodo de tipo comentario.
- **processing-instruction(),** devuelven cualquier tipo de instrucción de proceso.

2.8 SELECTORES

// -> todos los elementos igual a esa etiqueta que sean descendientes del elemento actual

Ejemplo: //BBB

http://zvon.org/xxl/XPathTutorial/Output_spa/example2.html

* -> selecciona todos los elementos

Ejemplo: /*/*/BBB

http://zvon.org/xxl/XPathTutorial/Output_spa/example3.html

text() -> selecciona sólo el texto de dicho elemento.

Ejemplo: /AAA/text()

2.9 FUNCIONES

boolean(), al aplicarla sobre un conjunto de nodos devuelve true si no es vacío.

not(), al aplicarla sobre un predicado devuelve true si el predicado es falso , y falso si es true.

true(), devuelve el valor true.

false(), devuelve el valor false.

count(), devuelve el número de nodos que forman un conjunto de nodos.

name(), devuelve un nombre de un nodo.

local-name(), devuelve el nombre del nodo actual o del primer nodo de un conjunto.

namespace-uri(), devuelve el URI del nodo actual o del primer nodo de un conjunto dado.

position(), devuelve la posición de un nodo en su contexto comenzando en 1. Por ejemplo, para seleccionar los dos primeros elementos: //elemento[position()<=2]

last(), Devuelve el último elemento del conjunto dado.

normalize-space(), Si una cadena tiene espacios consecutivos, lo sustituye por uno solo.

string(), es una función que convierte un objeto en una cadena. Por ejemplo, los valores booleanos se convierten en la cadena que representa su valor, esto es “true” o “false”.

concat(), concatena dos cadenas. El ejemplo siguiente devuelve “Xpath permite obtener datos de un fichero XML”: concat('XPath', 'permite obtener datos de un fichero XML')

string-length(), devuelve la cantidad de caracteres que forman una cadena de caracteres.

sum(), devuelve la suma de los valores numéricos de cada nodo en un conjunto de nodos determinado.

2.10 PREDICADOS

PREDICADO

Exigir a un nodo ciertas características

Patrón: Corchetes

Ejemplos:

/libro/capitulo[@num="1"]/parrafo

//libro[numero] -> Selecciona el libro que ocupa esa posición. //BBB[2]

//libro[last()] Selecciona el libro que ocupa la última posición

//libro[@cod] Selecciona el libro que tiene ese código

2.11 CONDICIONES

CONDICIONES

//libro[@cod="1"] Selecciona el libro cuyo código sea igual a 1

//libro[autor="Cervantes"] Selecciona el libro cuyo autor sea Cervantes

//autor[.="Cervantes"] Para hacer referencia al propio valor se utiliza el punto (.)

//libro[@CAT="P"]

//libro[PRECIO>10]

//libro[@CAT="P"]/AUTOR/text()

//libro()]

//libro[starts-with(TITULO, 'N')]

//libro[starts-with(TITULO, 'N')]/AUTOR/text()

//libro[string-length(TITULO)>10]/TITULO

//libro[not(position() = last())]

Ejemplos XPATH

Resumen: https://www.mclibre.org/consultar/xml/lecciones/xml-xpath-resumen.html

https://www.mclibre.org/consultar/xml/lecciones/xml-xpath.html

Trabajo clase 1. XPATH

Dado **Movies.xml** en la entrega de tarea, crea las siguientes consultas y entrégalas en un .txt:

1. Películas con review=”5”
2. Películas con review=”5” y del año 1992 (necesario operador and)
3. Películas cuyo año sea par (necesario operador mod)
4. Segunda película
5. Películas en las que ha intervenido el actor Nicolas Cage (text()=‘Nicolas Cage’)
6. Título de las películas en las que ha intervenido el actor Nicolas Cage
7. Actores que han trabajado con el actor Nicolas Cage en alguna película, incluido el propio Nicolas Cage
8. Actores que han trabajado con el actor Nicolas Cage en alguna película, excluido el propio Nicolas Cage (operador !=)
9. Productores que han producido películas del anio 1992
10. Título de las películas comedia interpretadas por Nicolas Cage (operador and)
11. Películas con tres o más actores (función count())
12. Titulo de las películas con tres actores
13. Titulo de las películas que tienen un productor apellidado Wood (función contains())
14. Última película (función last())
15. Todas las películas excepto la ultima (funciones not(), position() y last())
16. Todos los elementos que contengan el atributo year
17. Todas las películas que no contengan subelemento comments (función not()))
18. Todos los nodos actor o director. (uso de |)
19. Titulo de las películas donde participe algún miembro de la familia Coen (uso de función contains())