EJERCICIOS XPATH

1. Tomando el archivo **catalogo.xml** selecciona los siguientes elementos:

   a) Todos los nodos *producto*
   b) Todos los nodos *medium*
   c) Todos los atributos *imagen*
   d) Todos los nodos *precio*
   e) Todos los nodos *tallas*
   f) El texto de todos los nodos *precio*
   g) El texto de los nodos *id_modelo*
   h) Los nodos color de las tallas *small*
   i) Todos los nodos que desciendan de *producto*
   j) Todos los nodos que desciendan de *modelo*
   k) El texto de los nodos *opinion*
   l) Todos los nodos *opiniones*
   m) Todos los atributos *talla*
   n) Todos los textos
   o) Todos los atributos

2. Tomando el archivo **catalogo.xml** selecciona los siguientes elementos:

   a) Los nodos padres de aquellos que contengan el atributo *imagen*
   b) Los modelos de los que haya, al menos, una *opinion* escrita.
   c) Los nodos que tenga un atributo *talla* o un atributo *genero*
   d) Los nodos asociados a las tallas *small* o *xlarge*
   e) Los textos de los nodos de color asociados a las tallas *small* o *xlarge*
   f) Los atributos de los nodos de color asociados a las tallas *small* o *xlarge*

3. Tomando el archivo **cursos.xml** selecciona los siguientes elementos:

   a) Los cursos del instructor “Brightman”
   b) Los cursos cuya materia (*subj)* sea “BIOL”
   c) Los cursos que se impartan lunes, miércoles y viernes (M-W-F)
   d) Los cursos que se impartan lunes, miércoles y viernes (M-W-F) o bien los jueves (Th)
   e) Los cursos que se impartan lunes, miércoles y viernes (M-W-F) en el aula 116 (room)
   f) El título de los cursos que se impartan lunes, miércoles y viernes (M-W-F) en el aula 116 (room)
   g) El texto del título de los cursos que se impartan lunes, miércoles y viernes (M-W-F) en el aula 116 (room)
   h) Los cursos que no sean ni de química (CHEM), ni de biología (BIOL), ni de antropología (ANTH)
   i) El texto de las horas de comienzo y finalización de los cursos que no sean ni de química (CHEM), ni de biología (BIOL), ni de antropología (ANTH)

4. Tomando el archivo **cursos.xml** selecciona los siguientes elementos:

   a) El texto de los títulos de los cursos del instructor “Brightman”
   b) La hora de comienzo de los cursos del instructor “Brightman”
   c) El texto de los títulos de los cursos del instructor “Brightman” y la hora de comienzo de los cursos del instructor “Brightman” (pista: usa |)
   d) Los cursos que tengan el valor 0.0 en las unidades (*units)*
   e) La materia (subj) de los cursos que tengan el valor de *units* menor que 1.0
   f) La materia (subj) de los cursos que tengan el valor de *units* menor que 1.0 pero mayor que 0.0
   g) El texto del título de los cursos que tengan el valor de *units* a 1.0 y que sean de la asignatura de antropología (ANTH) y que se impartan en el aula 120 (room).
   h) El texto del aula en el que se imparten los cursos con más de 0.5 unidades, de la materia de economía (ECON) y que se impartan en lunes, miércoles, viernes (M-W-F)
   i) El texto de la hora de comienzo de los cursos con menos de 1.0 unidades y de la materia de biología (BIOL)
   j) El título de los cursos que se impartan en el edificio (building) de química (CHEM)

5. XPath también se puede usar para “web scraping” (conjunto de técnicas que permiten extraer información de la web de forma masiva). Para “coquetear” con esta técnica vamos a scrapear con XPath la siguiente web:

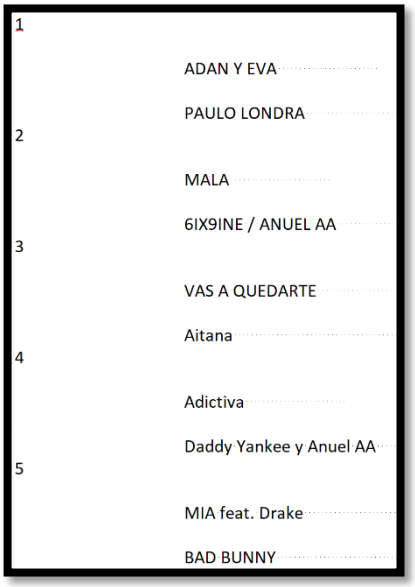
   https://www.elportaldemusica.es/lists/top-100-canciones/2019/1

   Para ello, debes descargar el fichero *“PortalMúsica.zip”* y descomprimirlo.

   El HTML ha sido previamente tratado del siguiente modo:
   - Se han desactivado los script de javascript para que se abra rápidamente en un navegador.
   - Se han “arreglado” todas las etiquetas que no cumplen con la sintaxis XML (se han cambiado las
 por
, las <img …> por)

   La idea es usar el XML Copy Editor para sacar a un fichero de texto la información de la imagen.

   Se recomienda hacer una expresión para sacar los números, otra para el nombre de las canciones y otra para el nombre del artista. Después se relacionan con un | OR.



6. Tomando el fichero **matricula-cursos.xml** vamos a realizar las siguientes selecciones:

a) Modifica el ejemplo estudiado en las diapositivas para extraer el número de las aulas en las que se imparten los cursos en los que se ha matriculado Thomas Fersen

b) En dos pasos:
   a) Selecciona los reg_num de los alumnos cuyo país sea España
   b) A continuación, muestra los días de la semana en los que se imparten los cursos de los alumnos cuyo país sea España.

c) En dos pasos:
   a) Selecciona los reg_num de los alumnos cuyo país sea España o Francia
   b) A continuación, muestra el nombre de los instructores de los cursos que hayan escogido el alumnado proveniente de España o Francia.

d) Muestra la materia (subj) de los cursos seleccionados por el alumnado que provenga de Rusia.

e) Muestra el horario de comienzo y de fin de los cursos matriculados por los alumnos John No Fear y Thomas Fersen.

f) Muestra el nombre de los cursos matriculados por el último alumno de la lista.

g) Muestra el horario de comienzo y de fin de los cursos matriculados por el primer alumno de la lista y el último.

h) Muestra el nombre de los alumnos que se han matriculado en el curso con título “Introduction to Combinatorics”

i) Muestra el país de procedencia de los alumnos que se han matriculado en el curso con título “Intro to Probability and Statist”

j) Muestra el nombre y el país de procedencia de los alumnos que se han matriculado en el curso con título “Introduction to Combinatorics”

k) Muestra el nombre de los alumnos que están matriculados en algún curso impartido por el instructor Shurman

l) Muestra el nombre de los alumnos que están matriculados en algún curso en el que haya algún alumno procedente de Alemania.

7. Vamos a utilizar este ejercicio para integrar/asentar todas las expresiones estudiadas con XPath, los apartados irán en complejidad creciente. Para ello tomaremos el fichero **spain.xml** realizaremos las siguientes selecciones:

   a) Todos los nodos de comunidades autónomas (*province)*
   b) El nombre de cada una de las comunidades autónomas
   c) El texto del nombre de cada una de las ciudades de España
   d) El texto de la población de cada una de las ciudades de España
   e) Todos los atributos de los nodos de las comunidades autónomas *(province)*
   f) Todos los atributos de los nodos de las ciudades
   g) Todos los atributos *province* de las ciudades
   h) Todos los nodos de población de ciudades y también todos los nodos de idiomas (languages)
   i) Todos los nodos de nombres de ciudades y también el nodo que contiene el nombre del país
   j) Todos los nodos de ciudad de la comunidad autónoma de “Castile La Mancha”
   k) El texto de todos los nodos de ciudad de la comunidad autónoma de “Galicia”
   l) El texto de todas las poblaciones que no pertenezcan a la comunidad de Madrid
   m) El texto de todas las ciudades que no pertenezcan a ningún archipiélago de islas
   n) La última comunidad autónoma del listado
   o) La penúltima ciudad de cada comunidad autónoma
   p) La primera ciudad de todas las comunidades autónomas
   q) La primera ciudad de la primera comunidad autónoma.
   r) El nodo de la ciudad asociada al atributo *capital* del nodo *country*.
   s) El nombre de los paises con los que España comparte frontera (*border)*
   t) La población de los paises con los que España comparte frontera
   u) El nombre de las capitales de cada comunidad autónoma.
   v) La población de las capitales de cada comunidad autónoma.
   w) El nodo de las comunidades autónomas que tengan un área menor que el área de Andalucía.
   x) El nombre de las comunidades autónomas que tengan una población mayor que la comunidad autónoma de Madrid.
   y) El nombre de las ciudades que tengan más población que Andorra.
   z) El nombre de los países que tengan más población que Portugal.