
Procesamiento de documentos XML

Ofimática Avanzada

Profesor: Víctor Fresno Fernández
curso 2006/07

DOM

Modelo de Objeto de Documento (DOM):

- Representación interna estándar de la estructura de un documento
- Proporciona un interfaz (API) al programador para acceder de forma fácil, consistente y homogénea a los elementos y atributos de un documento
- Es un modelo independiente de la plataforma y del lenguaje de programación

DOM

Un DOM completo debería permitir:

- Reconstrucción del documento completo a partir del modelo
- El acceso a cualquiera de las partes del documento
- Manipulaciones, adiciones y eliminaciones en el documento

El W3C se encarga de la especificación del DOM para HTML y XML

DOM

Hay 3 niveles de especificación del DOM:

- Nivel 1: modelos para HTML y XML.
 - Funcionalidades para la navegación y manipulación de documentos
 - Tiene 2 partes: el “*core*” referida a XML y la parte HTML
- Nivel 2: modelo de objetos e interfaz de acceso a las partes de un documento
- Nivel 3 (a nivel de recomendación): Permite el acceso a las DTD, hojas de estilo, espacios de nombres

DOM

DOM presenta los documentos como una jerarquía de objetos nodo

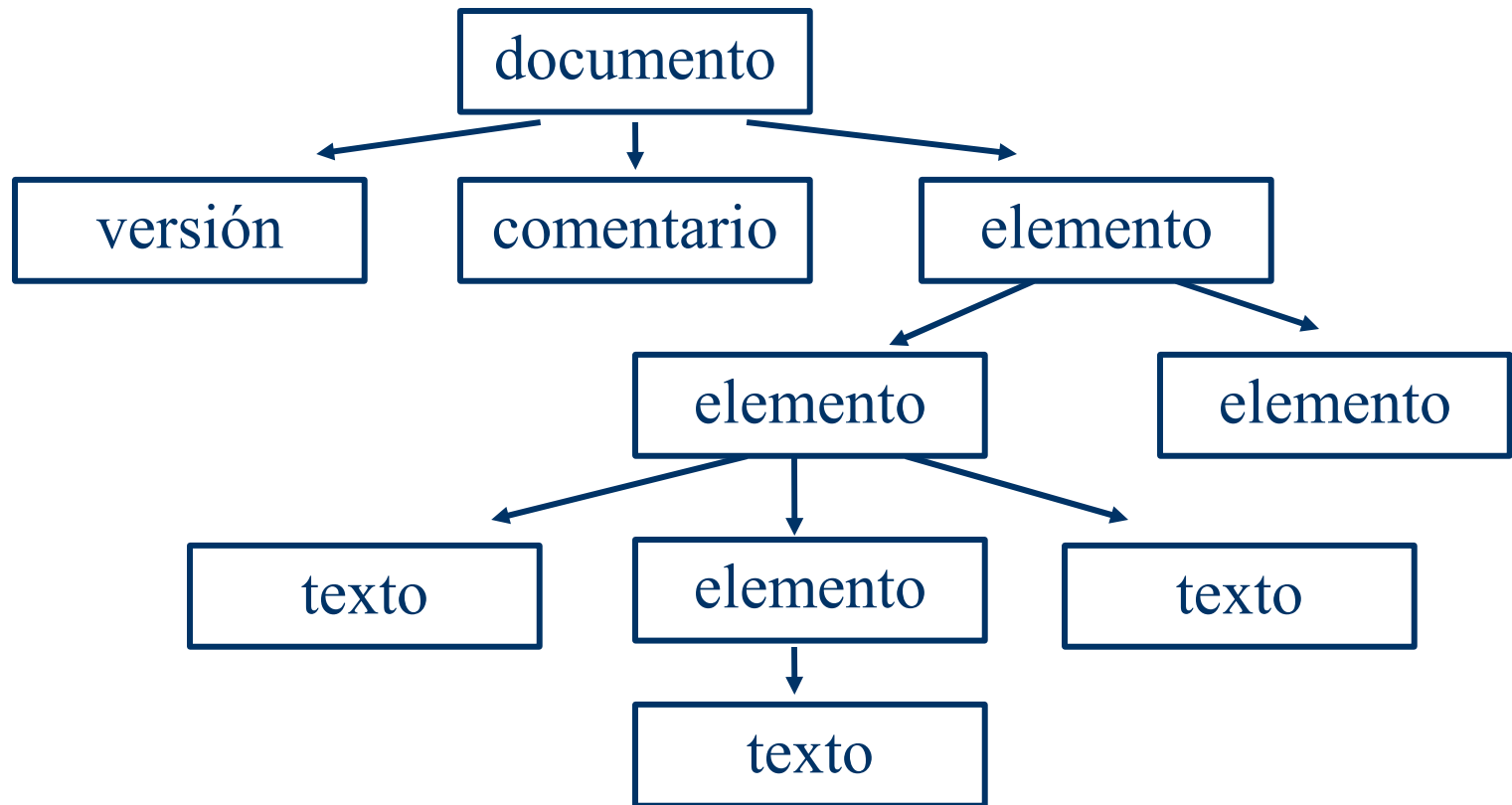
- Algunos tipos de nodos pueden tener nodos hijo
- Otros nodos son nodos hoja

DOM

Ejemplo:

```
<?xml version="1.0" encoding="UTF-16"?>
<!-- demostracion de una arbol de analisis-->
<doc>
  <saludo>Hola <enfaticado> parser </enfaticado>
  XML </saludo>
  <aplausos tipo="brutal"/>
</doc>
```

DOM



DOM

La especificación DOM del W3C utiliza el lenguaje OMG IDL (Object Management Group Interface Definition Language)

- En IDL:
 - Los objetos tienen interfaces con el mundo exterior
 - Cada interfaz tiene unos atributos que describen las propiedades del objeto
 - El objeto se manipula a través de unos métodos
 - Los métodos devuelven un resultado a la aplicación solicitante

DOM

- Existen diferentes implementaciones de DOM y SAX para distintos lenguajes de programación .

- (ADA95)

<http://www.nodix.de/xml4ada95/download.htm>

- (JAVA)

http://sourceforge.net/project/showfiles.php?group_id=16035

SAX

API sencilla para XML, **Simple API for XML (SAX)**:

- El analizador de SAX:
 - No crea ninguna estructura de datos para representar el documento
 - Va analizando el documento y generando eventos (comienzo de un elemento, final de un elemento, ...)
- SAX es una interfaz de un analizador más que una API para una estructura de datos en árbol (DOM)

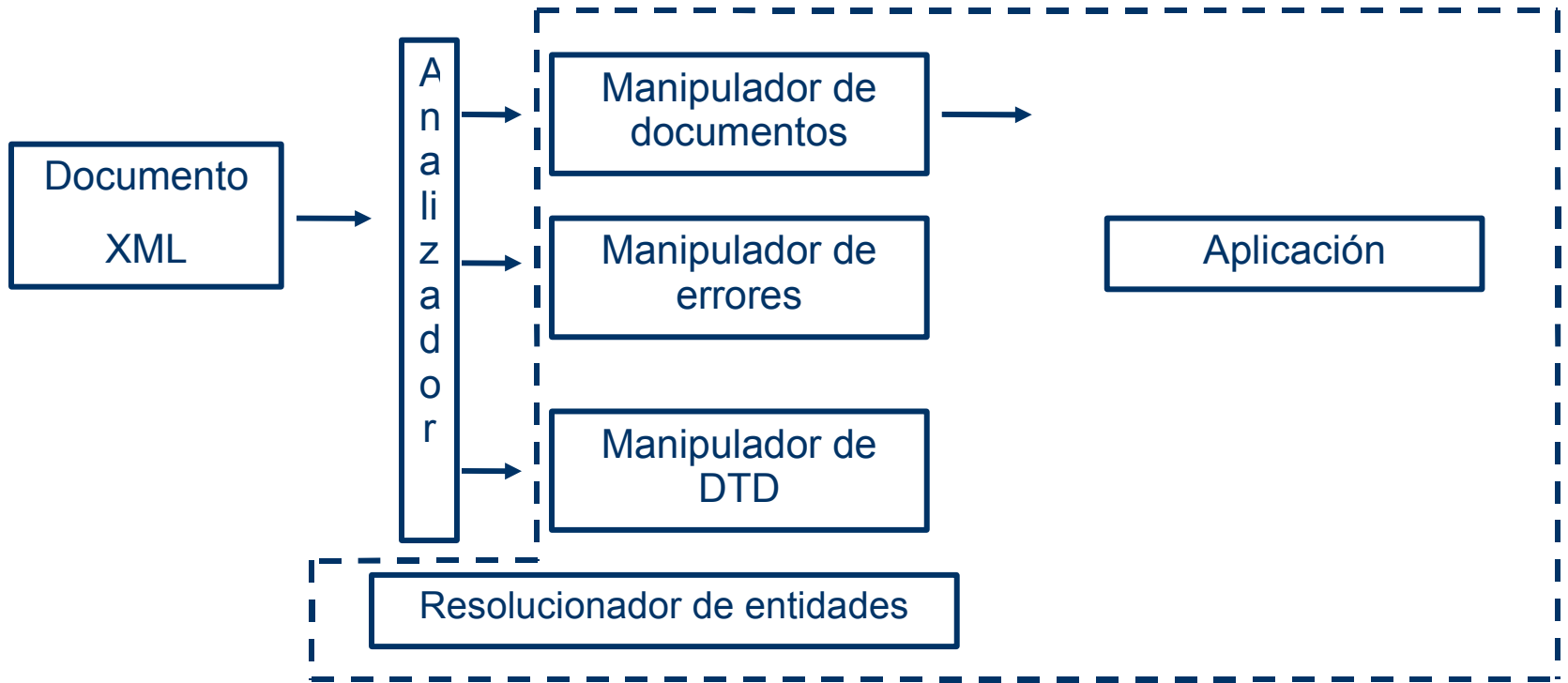
SAX

- Está en un nivel más bajo que DOM
- Será mejor que DOM cuando:
 - El documento no quepa en memoria
 - Las tareas sean irrelevantes para la estructura del documento (contar todos los elementos, extraer el contenido de un elemento específico, ..)

SAX

- La aplicación debe registrar manipuladores de eventos a un objeto analizador que implementa la clase `org.sax.Parser`
- SAX tiene 3 interfaces de manipuladores:
 - DocumentHandler, DTDHandler, ErrorHandler
 - DocumentHandler es la más importante
- Existen diferentes implementaciones de SAX para distintos lenguajes de programación .

SAX



El analizador empaqueta los datos XML en eventos que la aplicación puede manipular

