Descubrimiento de información en textos

Tarea del tema 3: Estudio de estándares de anotaciones

Gabriel Vázquez Torres

Descubrimiento de información en textos

Tarea del tema 3: Estudio de estándares de anotaciones

Contenido

Descripción de la práctica

A partir de un fichero en formato xml:

1. Comprobar si dicho documento XML es un documento conforme con la DTD de TEI Lite o con un subconjunto conforme con TEI Lite. Explicar por qué lo es o no.
2. En caso de que no lo sea, realizar las adiciones y sustituciones de elementos y atributos necesarias para que lo sea. Explica el porqué de cada adición y sustitución. Tened en cuenta que cuando un elemento o atributo no sea correcto para TEI Lite, habrá que intentar sustituirlo por otro que sí lo sea y que permita mantener la anotación.
3. Diseña una DTD que sólo contemple los elementos y atributos XML que hayas utilizado del TEI Lite.

Apartado 1

Analizando el documento XML y comparando sus etiquetas con las referencias dada en la teoría del tema 3 de la asignatura: http://www.tei-c.org/Guidelines/Customization/Lite/ y http://www.tei-c.org/Guidelines/Customization/Lite/ podemos concluir sin atisbo de duda que el documento XML **NO** está conforme con la DTD de TEI Lite. Aunque existen etiquetas que sí cumplen el DTD de TEI Lite como por ejemplo la etiqueta <p>, <list> o <date>, otras muchas no lo cumplen como son el caso, por ejemplo, de las etiquetas <corpus> o <ITEM>. Para más inri, nos encontramos con etiquetas como <encabezado>, <tipo>, <idioma> o <colon> que no son la manera correcta de etiquetar tanto una cabecera en TEI Lite como los dos puntos, como podemos ver en el ejemplo de la página http://lists.xml.org/archives/xml-dev/200003/msg00310.html . También, en las etiquetas <rs> los atributos "id" no están escritos correctamente, las comillas siempre tienen que ser dobles para los valores de los atributos y etiquetas como <seg#9> o <seg#10> no están bien nombradas. [1]

Apartado 2

Las sustituciones que he realizado sobre el fichero de ejemplo han sido las siguientes:

1. A la hora de asignar un id a los elementos de tipo <rs>, se ha modificado el atributo id por el atributo key de la siguiente manera: <rs type=”law” id=”LES2”> por <rs type=”law” key=”LES2”>
2. La etiqueta <ITEM> ha sido sustituido por la etiqueta <TEI.2>.
3. La etiqueta <tipo> ha sido sustituido por el atributo type en la etiqueta <teiHeader>. Ej: <teiHeader type="x">, siendo x un valor cualquiera y coherente con el texto.
4. La etiqueta <idioma> ha sido sustituido por el atributo lang en el elemento <teiHeader>. E.g: <teiHeader type="x" lang="es">, siendo x cualquier palabra coherente con el texto.
5. La etiqueta <corpus> ha sido sustituido por la etiqueta <teiCorpus>.
6. En las etiquetas <item> se ha modificado n=X por n=”X”, es decir, cumpliendo con el TEI Lite, todos los valores han sido puestos entre comillas.
7. Todos los atributos que tenían un valor asignado sin unas comillas previas, han sido modificados con comillas, por ejemplo: <rs type=law id=LES2> se ha modificado a <rs type=”law” id=”LES2”>. En estos casos, por ejemplo, se ha tenido que seguir el patrón de la página web oficial para esa etiqueta, en este caso de rs: http://www.tei-c.org/release/doc/tei-p5-doc/es/html/ref-rs.html. De esta forma hemos podido determinar qué elementos podía contener o no la etiqueta.
8. En las etiquetas <seg> se han modificado su manera de numerarlos. Las “#”, se han eliminado. Por ejemplo: <seg#10> → <seg10>. En este caso, aunque en la página oficial no se distinga en los ejemplos (http://www.tei-c.org/release/doc/tei-p5-doc/es/html/ref-seg.html), se ha comprobado que el elemento “seg” permite numeración.
9. La etiqueta <encabezado> ha sido sustituido por la etiqueta <teiHeader>.
10. Las etiquetas <colon> y </colon> se han eliminado.

Todas estas sustituciones han sido modificadas sabiendo que las etiquetas deben estar en la página http://www.tei-c.org/release/doc/tei-p5-doc/es/html/REF-ELEMENTS.html, siguiendo el DTD de la página http://www.tei-c.org/release/xml/tei/custom/schema/dtd/tei_lite.dtd y fijándonos en el ejemplo de la página http://lists.xml.org/archives/xml-dev/200003/msg00310.html como referencias.

- Como se ha dicho antes, hay muchas otras etiquetas que no se han tocado, como son los caso de <text>,<body>, <div1>,<div2> (como curiosidad, en la página oficial solo se ofrece hasta el “div7”), <p>, o <s>.

Apartado 3

Para este apartado se ha tenido en cuenta todas las anotaciones presentes en la documentación oficial y principalmente la estructura y contenido del siguiente documento web: http://www.tei-c.org/release/xml/tei/custom/schema/dtd/tei_lite.dtd .

En este caso, al utilizar el programa XML Copy Editor, he comprobado que existe la opción de crear un Schema (DTD) directamente desde el programa, pero no he conseguido hacerlo automáticamente por lo que he tenido que realizarlo de forma manual a partir de ejemplos como la página anterior.

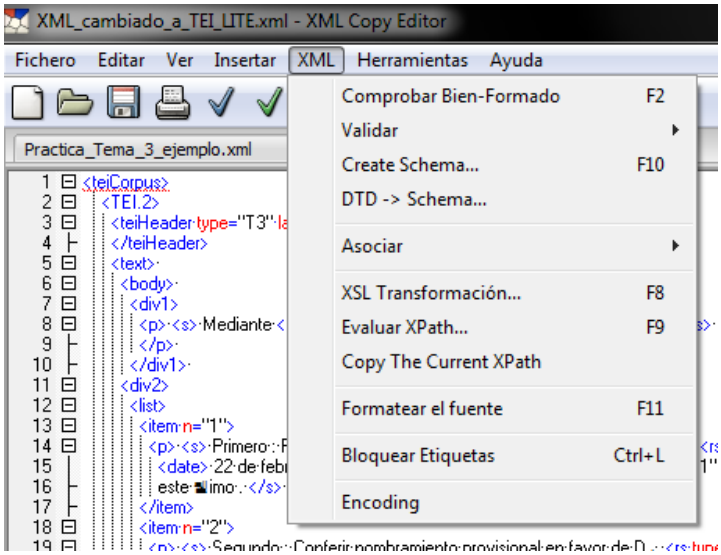

Ilustración 1: XML--> Create Schema

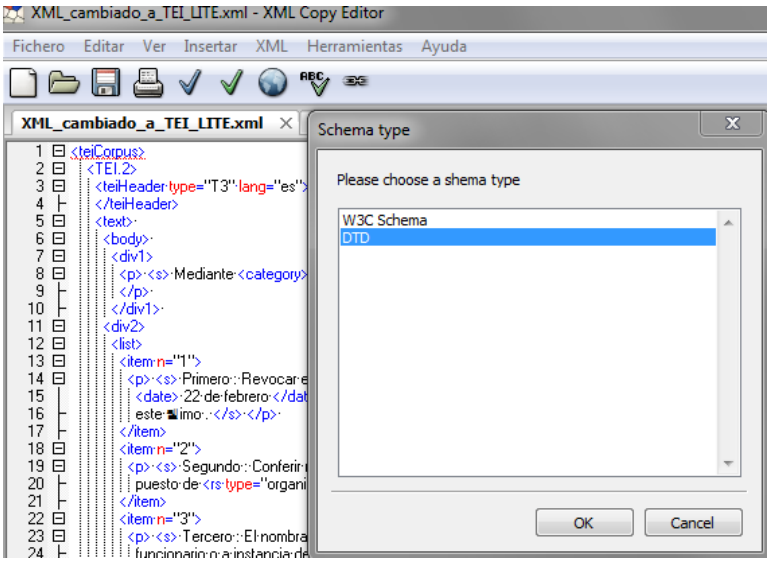

Ilustración 2: DTD

Al seleccionar la opción, nos ha aparecido una alerta, por lo que hemos decidido usar la forma manual.

Bibliografía

[1] [En línea]. Available: http://www.tei-c.org/release/doc/tei-p5-exemplars/pdf/tei_lite.doc.pdf.

Todas las bibliografías se han ido añadiendo a lo largo del documento por lo que están implícitas en dicho documento y no en la bibliografía.