Desarrollo de herramientas informáticas para el marcado multilingüe: conversión del TEI al español

Alejandro Bia y Manuel Sánchez Quero

Biblioteca Virtual Miguel de Cervantes,
Universidad de Alicante,
Apdo. de correos 99, E-03080, Alicante, España
alexbia, manuel.Sánchez@cervantesvirtual.com
http://cervantesvirtual.com/

Resumen. En este artículo mostraremos y defenderemos los beneficios del uso de esquemas de marcado XML multilingües para grandes proyectos de digitalización como la Biblioteca Virtual Miguel de Cervantes[1] y el consecuente incremento en la producción debido a la utilización de las etiquetas en la lengua propia. Del mismo modo, también mostraremos el proceso que se ha llevado a cabo para diseñar el método y desarrollar los programas necesarios para la generación automática de transformaciones XSL (*Extensible Stylesheet Language*) y programas en C que convierten el marcado en inglés de los documentos XML, DTDs (*Document Type Definition*) y/o Schemas al español y a otros idiomas. Finalmente, mostraremos las conclusiones extraídas de la implementación de este proyecto así como los resultados del uso de marcado XML en español en nuestra biblioteca digital.

Keywords: Creación de documentos, DTD, Edición electrónica, Marcado, Schemas, TEI, XML, XSLT.

1 Introducción

La Biblioteca Virtual Miguel de Cervantes (BVMC) es uno de los mayores proyectos de edición y publicación electrónica en España, y tal vez la biblioteca digital de textos en lenguas iberoamericanas más grande de la Web, actualmente con más de 10.000 entradas en su catálogo [4]. La BVMC produce una media de 150 textos digitales en XML (*Extensible Markup Language*) al mes, la mayoría de los cuales son clásicos españoles desde el siglo XII hasta la actualidad, englobando una amplia variedad de temas y estilos literarios tales como poesía, narrativa, teatro, historia, geografía, derecho, etc. Estos textos

[1] http://cervantesvirtual.com

están dirigidos tanto a los lectores aficionados como a los investigadores especializados que aprovechan las posibilidades del marcado estructural complejo para su investigación.

En un proyecto como este, la cantidad y la calidad de la producción están asociadas en gran medida a la tecnología. Pequeños cambios en ciertos aspectos críticos de la tecnología de producción producen cambios considerables en los tiempos, en los costes y en la calidad del producto final. Hemos llevado a cabo, durante los cuatro años de vida que tiene actualmente la BVMC, varios proyectos para mejorar el proceso de producción de los libros digitales en XML, tales como la construcción de herramientas especializadas para la corrección ortográfica de castellano antiguo de diferentes épocas [5], el diseño de procedimientos y herramientas para la simplificación automática de DTDs [3][2]. Hemos desarrollado programas para convertir ficheros de otros formatos tales como RTF y HTML a XML, y de XML a otros formatos, y también software para controlar el flujo de producción (*production workflow*) y la gestión documental, así como algoritmos para la estimación de costes de producción [1].

2 Marcado, significado y multilingüismo

En 1998 Robin Cover escribió: "¿De qué modo ayuda el XML al etiquetado de información a nivel semántico? Los nuevos usuarios a veces se refieren al XML como un marcado semántico, y se puede oír cómo alaban el XML por su habilidad para expresar claridad semántica por medio del marcado... Alguien que use un editor de textos para examinar un documento XML – comparándolo con un archivo antiguo WordStar, con un fichero de texto delimitado por comas, con Postscript, o con cualquier documento que use un lenguaje de marcado procedural o presentacional – verá rápidamente que el documento XML tiene más significado respecto a los objetos de información representados por medio de texto. El marcado en sí mismo es un tipo de 'metadatos', que nos explica cuáles son los elementos constitutivos, y cómo estos objetos de información están estructurados en unidades mayores" [6].

A pesar de preferir actualmente la lógica de predicados como una alternativa más adecuada, para propósitos semánticos, a las DTDs convencionales [7], Sperberg-McQueen et al. defienden la utilidad del marcado como fuente de significado: "La función del marcado no es casual. El marcado tiene significado. ¿Qué significa tener significado? ¿Cómo es que tiene significado el marcado? ¿Por qué preocuparse por este asunto?: Simplemente para lograr una mejor documentación del lenguaje de marcado, para un mejor control de la calidad, para tener mejores procesos automáticos (traducción, normalización, consulta), para ofrecer un modo de examinar las prácticas habituales... y porque es interesante. ¿Cómo es que el marcado tiene significado? Porque el marcado significa algo... sabemos ciertas cosas. Por ejemplo, debido a que vemos cierto marcado,

podemos hacer ciertas inferencias" y concluyen que "el significado del marcado es el conjunto de inferencias que este permite hacer" [9].

Por tanto, uno de los aspectos clave del marcado estructural es el significado que este tiene, que depende de nuestra habilidad para entenderlo. Entender las etiquetas XML es básico para delimitar correctamente estructuras textuales complejas para su posterior procesamiento automático. Este entendimiento puede verse perjudicado cuando los nombres de las etiquetas (elementos, atributos y los valores de los atributos) están en otro idioma.

Nuestra biblioteca digital es un proyecto multidisciplinar donde especialistas de diferentes áreas de estudio (filólogos, informáticos, bibliotecarios, sociólogos, etc.) trabajan conjuntamente. El grupo más grande, por lejos, de especialistas de la biblioteca lo constituye el área de corrección y marcado (50 personas aproximadamente), compuesto por especialistas de diferentes campos de las humanidades, ninguno de ellos relacionado con la lengua inglesa. Es en este área donde la necesidad e importancia de traducir el marcado original del inglés al español se hace más evidente.

Gracias a la práctica hemos aprendido a utilizar un conjunto de etiquetas en una lengua ajena, el inglés, pero el uso de etiquetas de marcado en otro idioma aumenta el tiempo de aprendizaje y reduce la calidad y la cantidad de textos digitales producidos. Los nombres de las etiquetas son mnemotécnicos que pueden sonar familiares para los hablantes de inglés pero que son difíciles de entender y memorizar para los usuarios de otras lenguas. Al dar a nuestros etiquetadores la posibilidad de utilizar etiquetas en español se ha incrementado la cantidad y calidad de la producción de textos digitales.

Por ejemplo, utilizar etiquetas del tipo `<???????? ???????="???">` hace que se pierda todo el significado en el proceso de marcado de documentos. Así pues, imaginemos el caso de tener que utilizar un esquema de marcado en un idioma como el ruso, donde el nombre de los elementos, atributos y valores de atributos no tienen ningún valor semántico para etiquetadores que no dominan esta lengua. Ahora bien, todo cambia si en lugar de utilizar la etiqueta anterior utilizásemos su equivalente en español: `<titulo tipo="principal">`.

Dado que estamos totalmente convencidos del valor y las ventajas que conlleva el uso de estándares hemos elegido el vocabulario de marcado TEI (*Text Encoding Initiative*) que es un estándar de facto, al menos entre la comunidad de investigadores y estudiosos de literatura inglesa. Tras usarlo con éxito durante algún tiempo en inglés, nos embarcamos en el proyecto de traducir los nombres de los elementos, atributos y valores del TEI al español. Finalmente, desarrollamos las herramientas de traducción para garantizar la conversión automática de y al conjunto de etiquetas del TEI en inglés. Estos programas de conversión automática traducen no sólo el marcado de documentos XML sino también de las correspondientes DTDs.

Ahora estamos en proceso de construir otros conjuntos de etiquetas del TEI y traducciones a diferentes idiomas. El objetivo es tener muchas traducciones oficiales del TEI, pero solamente una versión principal (la original en inglés). La automatización de la traducción de las etiquetas a otras lenguas es vital para asegurar un intercambio sencillo de documentos entre proyectos que utilizan diversas lenguas. De este modo, y desde un punto de vista estructural y semántico, el conjunto de etiquetas es el mismo, simplemente cambia el nombre.

También consideramos que tener versiones multilingües de un conjunto de etiquetas dado, como el TEI, puede facilitar su introducción en muchas partes del mundo como Sudamérica, donde el uso del XML para la edición electrónica es todavía bastante inusual. Esto puede resultar de especial interés para las bibliotecas y editoriales digitales de todo el mundo.

La principal razón para llevar a cabo esta iniciativa es que los esquemas de marcado normalmente están definidos en inglés y existe una gran comunidad de usuarios que no utilizan el inglés de forma tan fluida y por tanto pierden su significado. Si el esquema de marcado se traduce a la lengua de los usuarios, el proceso de asimilación y control de su utilización se aceleraría y la producción de textos marcados aumentaría, con la consecuente reducción de costes.

3 Generación automática de traductores de marcado

Comenzamos definiendo un conjunto de posibles traducciones de nombres de elementos, atributos y valores de atributos para los diferentes idiomas de destino. Almacenamos dicha información en un documento XML de mapeo para la traducción multilingüe. El siguiente es un ejemplo de este documento y de su DTD.

DOCUMENTO DE MAPEO PARA LA TRADUCCIÓN DE INGLÉS, ESPAÑOL Y FRANCÉS:

```
<TAGMAP>
...
  <ELEMENT en="body" sp="cuerpo" fr="corps">
  </ELEMENT>
...
  <ELEMENT en="div0" sp="div0" fr="div0">
     <ATTR en="lang" sp="lengua" fr="langue">
     </ATTR>
     <ATTR en="type" sp="tipo" fr="type">
```

```
       <VALUE en="news" sp="noticias" fr="nouvelles"/>
       <VALUE en="suggestions" sp="sugerencias"
fr="sugestions"/>
       <VALUE en="biblnews" sp="novedades"
fr="publications"/>
     </ATTR>
  </ELEMENT>
...
  <ELEMENT en="p" sp="parrafo" fr="paragraphe">
     <ATTR en="align" sp="alinear" fr="aligne">
       <VALUE en="left" sp="izq" fr="gauche"/>
       <VALUE en="right" sp="der" fr="droite"/>
       <VALUE en="center" sp="centro" fr="centre"/>
       <VALUE en="justify" sp="justificar" fr="justifie"/>
     </ATTR>
     <ATTR en="indent" sp="sangria" fr="retraitpositif">
       <VALUE en="left" sp="izq" fr="gauche"/>
       <VALUE en="right" sp="der" fr="droite"/>
       <VALUE en="both" sp="ambas" fr="lesDeux"/>
       <VALUE en="none" sp="ninguna" fr="aucune"/>
     </ATTR>
     <ATTR en="specialindent" sp="sangriaespecial"
fr="retraitnegatif">
       <VALUE en="none" sp="ninguna" fr="aucune"/>
       <VALUE en="firstline" sp="primeralinea"
fr="premiereLigne"/>
       <VALUE en="french" sp="francesa" fr="francaise"/>
     </ATTR>
  </ELEMENT>
...
</TAGMAP>
```

DTD DEL ARCHIVO ANTERIOR:

```
<!ELEMENT TAGMAP (ELEMENT)+ >

<!ELEMENT ELEMENT (ATTR)* >
<!ATTLIST ELEMENT
     en CDATA #REQUIRED
     sp CDATA #REQUIRED
     fr CDATA #REQUIRED>

<!ELEMENT ATTR (VALUE)* >
```

```
<!ATTLIST ATTR
     en CDATA #REQUIRED
     sp CDATA #REQUIRED
     fr CDATA #REQUIRED>

<!ELEMENT VALUE EMPTY >
<!ATTLIST VALUE
     en CDATA #REQUIRED
     sp CDATA #REQUIRED
     fr CDATA #REQUIRED>
```

Este documento de mapeo que contiene toda la información necesaria estructurada para desarrollar los conversores de idioma lo lee el generador de transformaciones, que se construyó en forma de una hoja de estilo XSLT [8]. El XSL puede utilizarse para procesar documentos XML para producir bien otros documentos XML o bien un documento de texto plano. Dado que las hojas de estilo XSL son XML, pueden generarse como salida otra hoja de estilo XSL. De este modo, y para cada uno de los idiomas incluidos en el archivo de mapeo para la traducción multilingüe, producimos una transformación del inglés a la lengua local así como otra transformación de la lengua local al inglés. De esta forma aseguramos la conversión de los documentos XML en ambas direcciones.

También generamos para cada idioma un traductor de DTDs en forma de programa escrito en C++ y Lex. Debemos tener en cuenta que las DTDs no cumplen la norma del XML y por tanto no pueden ser transformados usando XSLTs. Debido a esto hemos usado la capacidad del XSL para producir texto plano, ahora como un programa en C++. Solamente hemos considerado una traducción unidireccional de la DTD en inglés a la DTD en el idioma local, ya que asumimos que la DTD se construiría desde el principio en el idioma original del vocabulario de XML (inglés) y luego traducida a la lengua local, y no al revés. No vimos necesidad de traducir la DTD en la lengua local de vuelta al inglés (flecha discontinua), pero esta es una transformación que podría realizarse fácilmente del mismo modo si fuera necesario. Como la DTD en inglés puede usarse para validar el conjunto de documentos XML marcados en inglés, la DTD en lengua local puede utilizarse para validar el conjunto de ficheros marcados en la lengua local. Recientemente hemos agregado la capacidad de traducir esquemas (W3C Schemas), lo cual se hace mediante transformaciones XSLT, ya que los esquemas a diferencia de las DTD son XML.

Este proceso de generación de transformaciones se muestra para el español en la Fig. 1. Pueden construirse muchos otros traductores de marcado para otros idiomas del mismo modo. En nuestras pruebas hemos jugado con el inglés, español y francés, posibilitando la generación de transformaciones para traducir de uno a otro de estos idiomas, aunque la idea es traducir de y al idioma original del conjunto de etiquetas (inglés), que debe tomarse como lenguaje estándar para la transferencia de ficheros entre proyectos.

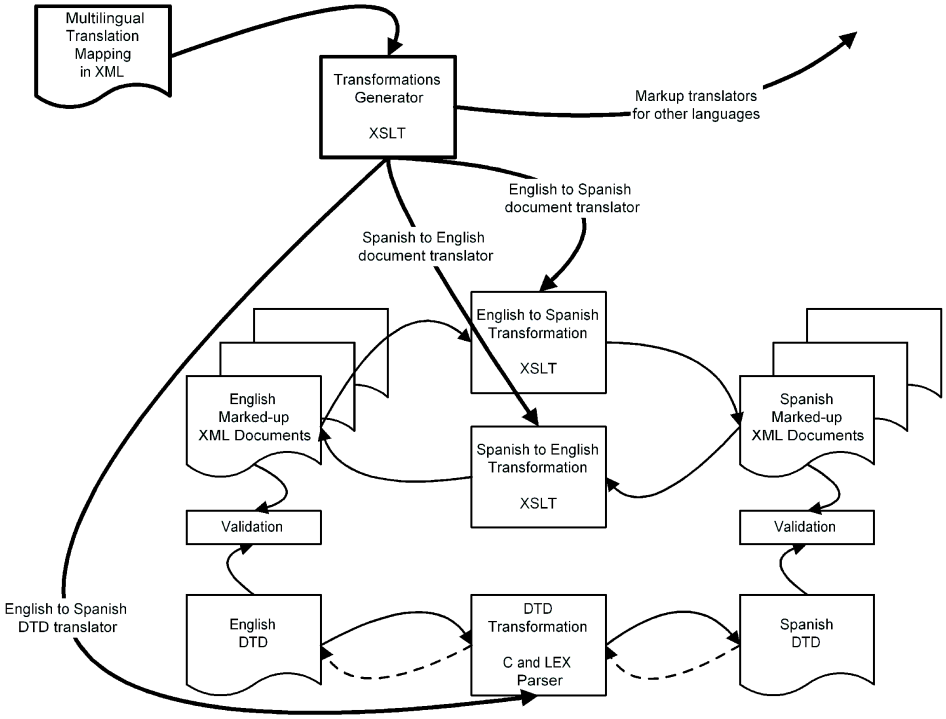

Fig. 1. Generación automática de traductores de marcado. Esta imagen muestra la generación de transformaciones XSL y programas C++ para convertir el marcado y las DTDs del inglés al español.

3.1 ¿Son los valores de los atributos dependientes del contexto?

A la hora de crear el documento XML de mapeo para la traducción multilingüe nos asaltó la duda de si un determinado valor de un atributo era traducible siempre por el mismo término. Obviamente, la respuesta a esta cuestión es que no, que los valores de los atributos son dependientes del contexto, esto es, que dependen directamente de a qué atributo y de a qué elemento están asociados.

Por ejemplo, puede darse el caso de una DTD donde tengamos los valores “SAT”, “SUN” y “MAR”. A primera vista estos podrían ser atributos referidos a fechas escritos en inglés, esto es: “SAT” correspondería a *Saturday* (sábado), “SUN” a *Sunday* (domingo) y “MAR” a *March* (marzo). No obstante, estos atributos pueden ser ambiguos ya que, dependiendo de en qué contexto se utilicen pueden hacer referencia a fechas o a

elementos del sistema solar: “SAT” por Saturno, “SUN” por el Sol (*Sun* en inglés) y “MAR” por Marte (*Mars* en inglés).

Así pues, queda claro que el contenido de los valores de atributos es interpretable según el contexto en el que se sitúa, es decir, que los valores de atributos dependen directamente del elemento o del atributo que los contienen.

3.2 Y los atributos... ¿son también dependientes del contexto?

Del mismo modo que pueden haber valores de atributos iguales para diferentes atributos con diferente interpretación, los atributos también pueden tener cierta ambigüedad en función de a qué elemento pertenecen. Aunque no es una idea muy acertada tener el mismo atributo en diferentes elementos con interpretaciones diferentes, no hay ninguna regla en XML que nos lo impida y es una opción a tener en cuenta. Así pues, podemos poner por caso la existencia de un atributo “w” para dos elementos diferentes (`<font>` y `<figure>`). En el caso del elemento `<font>` significaría *weight* y se emplearía para indicar si el texto va en negrita (*bold* en inglés) o no, mientras que el atributo “w” en el elemento `<figure>` se interpretaría como *width* para indicar la anchura de la imagen.

4 Explicación de uso

Desde un punto de vista minimalista, consideramos que el marcado en lengua local debería usarse casi de forma exclusiva para la creación y mantenimiento (p. ej., para etiquetar, editar y corregir los documentos XML). Existen otros casos donde también debería usarse el marcado en lengua local, como por ejemplo las búsquedas en XML cuando la interfaz del motor de búsqueda permite al usuario realizar consultas usando elementos, atributos y valores de atributos; en este caso hay una mayor comprensión si se realiza en la lengua local. Si el usuario tiene que usar los nombres de las etiquetas en otro idioma la ventaja semántica se pierde. Una interfaz de traducción también puede usarse para los motores de búsqueda en XML multilingüe.

Para el procesamiento automático y el intercambio de documentos creemos que en muchos casos es más conveniente usar el marcado en el idioma original del estándar de marcado (generalmente inglés). De este modo, las hojas de estilo y otras herramientas preexistentes, que han sido diseñadas para documentos marcados con marcado en el idioma original no necesitarán ser modificadas para aceptar textos con marcado en el idioma local, sino que en su lugar podrá traducirse el documento XML al idioma original de forma automática antes de ser procesado.

Este enfoque puede ser rebatido y de hecho puede haber usuarios que prefieran traducirlo todo (incluidas las hojas de estilo) al idioma local. Aunque no es estrictamente necesario, puede realizarse de forma sencilla.

Un riesgo que debemos evitar, si queremos conservar las ventajas de usar un esquema de marcado estándar ampliamente aceptado, es el desarrollo de esquemas de marcado alternativos en diferentes idiomas que pueden evolucionar o cambiar de forma independiente con respecto al estándar original.

Otro riesgo a evitar es la existencia de duplicados de un mismo documento, un problema muy conocido por los informáticos que puede tener muy serias consecuencias. Creemos que sólo debe existir un único documento fuente [4], sin importar el idioma que escojamos para su marcado, y todas las demás copias, de existir alguna, deben ser consideradas duplicados descartables o ficheros temporales. De este modo, el mantenimiento y modificaciones debería hacerse sobre el único documento fuente. Este es un problema de control de versiones que escapa a los objetivos de este proyecto, pero es un problema que merece mucha atención por parte de quienes elaboren documentos usando marcado en varios idiomas.

5 Trabajos previos

No nos ha sido fácil encontrar trabajos previos sobre marcado multilingüe. Sin dudas, habrá habido muchos esfuerzos aislados por traducir esquemas de marcado a otros idiomas para resolver problemas locales puntuales específicos, como afirma Wu [10], pero poco o nada de esto se ha publicado.

Tampoco tenemos conocimiento de vocabularios de marcado de uso generalizado que se hayan traducido a otros idiomas, menos aún que sean multilingües y que el proceso de traducción de sus elementos de marcado se haya automatizado. Sin embargo, hemos encontrado un trabajo interesante de Pei-Chi Wu [10], que trata el problema de la traducción de un vocabulario de marcado a otro idioma (de inglés a chino) para facilitar la comprensión y obtener un marcado más preciso. Este artículo trata el problema del marcado multilingüe, propone un procedimiento de traducción bilingüe y discute las aplicaciones potenciales de esto para el comercio electrónico. Describe un prototipo hecho en Java y MSXML (Microsoft XML) para el proceso de traducción de etiquetas, que se basa en DTDs paralelas equivalentes (una para cada lenguaje del vocabulario de marcado), en lugar de usar un fichero XML de traducción predefinido como se presenta en este artículo. En su proceso, primero construyen el fichero de traducción en tiempo de ejecución por comparación de las dos DTDs (la de origen y la de destino), y luego procesan el documento cambiándole las etiquetas. Comparándolo con el método que presentamos aquí, su método propone que el fichero de traducción se construya cada vez

que se transforma un documento (a pesar de que su prototipo no lo haga), no traducen ni los nombres de atributos ni los valores de los atributos cuando son por defecto, no generan automáticamente los programas de traducción de documentos ni abordan la traducción automática de las DTDs o de los Schemas. Sin embargo, su trabajo es un interesante antecedente que destaca la utilidad de este tipo de herramientas de traducción del marcado, al menos para el ámbito del comercio electrónico.

6 Conclusiones

- Los tiempos de aprendizaje se redujeron considerablemente.

- Los tiempos de producción también se redujeron, junto con un aumento en la calidad de marcado. Los etiquetadores se mostraron satisfechos y más seguros en su trabajo.

- Al utilizar el marcado en la propia lengua, el significado de las etiquetas y del marcado en general no se pierde.

- Proyectos multilingües pueden beneficiarse de la posibilidad de traducir de forma sencilla el marcado a la lengua de cada etiquetador.

- A menudo se desarrollan nuevos vocabularios de marcado no estándares simplemente porque resulta más sencillo que aprender un vocabulario estándar en un lengua extranjera. Al tener la posibilidad de usar un vocabulario estándar en la lengua propia se evita el desarrollo de nuevos vocabularios para satisfacer las necesidades de marcado locales. Esto puede contribuir a la difusión de los vocabularios de XML como TEI o DocBook en los países de habla no inglesa.

- La difusión del uso de vocabularios de marcado estándares es beneficioso para el intercambio de documentos.

7 Trabajo futuro

Este juego de herramientas puede ser rediseñado en un lenguaje de programación como Java o C++ para proporcionar un resultado más rápido y una interfaz más agradable. También puede ser implementado como un servicio Web.

Bibliografía

1. Alejandro Bia, DiCoMo: A cost estimation model for digitization projects, en ACH/ALLC 2002: New Directions in Humanities Computing, The 14th Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities, , Universidad de Tuebingen, Alemania (2002) 11-15.

2. Alejandro Bia y Rafael Carrasco: Generation of Simplified DTDs From a Set of XML Sample Files, en XML Europe 2002 Conference and Exposition, Hotel Princesa Sofía Inter-Continental, Barcelona, España (2002) p. 80. http://www.xmleurope.com/

3. Alejandro Bia, Rafael C. Carrasco y Manuel Sánchez Quero: A Markup Simplification Model to Boost Productivity of XML Documents, en Digital Resources for the Humanities 2002 Conference, Universidad de Edinburgo, George Square, Edinburgo EH8 9LD - Escocia - UK (2002) 13-16.

4. Alejandro Bia y Andrés Pedreño: The Miguel de Cervantes Digital Library: the Hispanic voice on the Web, en LLC (Literary and Linguistic Computing) journal, vol. 16, n. 2. Oxford University Press (2001) 161-177.

5. Alejandro Bia y Manuel Sánchez Quero: Building ancient Spanish dictionaries for spell-checking of DL texts, en LREC 2002, en Third International Conference on Language Resources and Evaluation (Manuel González Rodríguez y Carmen Paz Suárez Araujo, eds.), vol. VI. Las Palmas de Gran Canaria, España (2002) 1832-1837.

6. Robin Cover: Cover Pages XML and Semantic Transparency. Revisada el 24 de noviembre de 1998. (1998). http://www.oasis-open.org/cover/xmlAndSemantics.html

7. David Dubin, Michael Sperberg-McQueen, Allen Renear y Claus Huitfeldt: A logic programming environment for document semantics and inference, en ACH/ALLC 2002: New Directions in Humanities Computing. The 14th Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities, Universidad de Tuebingen, Alemania (2002).

8. Michael Kay: XSLT Programmer's Reference, Wrox Press, 1102 Warwick Road, Acocks Green, Birmingham, B27 6BH, UK, 1ª. ed. (2000).

9. C. M. Sperberg-McQueen, Claus Huitfeldt y Allen Renear: Meaning and Interpretation of Markup not as simple as you think, en Extreme Markup Languages, Montreal (2000).

10. Pei-Chi Wu (2000): “Translation of Multilingual Markup in XML”, 2000 International Conference on the theories and practices of Electronic Commerce, Part II, Session 14, pages 21-36, Association of Taiwan Electronic Commerce, Taipei, Taiwan, October 2000. URL: http://www.atec.org.tw/ec2000/PDF/14.2.PDF