Conversión de trabajos con contenido matemático a lenguaje HTML

Luis A. Piovan

Dedicado a la memoria de Eitan Gurari[1]

RESUMEN: Se discutirán herramientas usadas en el procesamiento de trabajos con contenido matemático para llevarlos a formato HTML.

1. TeX/LaTeX

El procesamiento de textos con contenido matemático para trabajos tipográficos de alta calidad ha estado dominado durante más de dos décadas por procesadores que se basan en el "markup language" (lenguaje de marcado). En los procesadores de texto usuales (Microsoft Word, OpenOffice, Apple Pages, NeoOffice, AbiWord, etc.) se puede ver el documento en la pantalla de la computadora mas o menos en la forma final que tendrá cuando impreso (WYSIWYG). El trabajo con un lenguaje de marcado (como el HTML o el LaTeX) es diferente en el sentido de que uno introduce lineas de comando de texto propias del lenguaje para producir un documento final con determinadas características. LaTeX es un lenguaje de marcado para la disposición de documentos diseñado e implementado por Leslie Lamport (1994) en A document Preparation System, basado en The TeXbook de Donald E. Knuth (1984). TeX, LaTeX y todas sus fuentes tipográficas son mantenidos por una creciente comunidad de desarrolladores esencialmente congregados en TeX Users Group.

La gran mayoría de las publicaciones científicas en Matemática, Física y Computación usan alguna implementación de LaTeX para su producción. Una de las razones para esto es que la fuentes tipográficas usadas por sistemas LaTeX (fuentes Computer Modern, LaTeX, AMS y otras) son especialmente diseñadas para exhibir las fórmulas matemáticas (al contrario de las fuentes propias de los procesadores de texto usuales). La geometría y arreglo del documento es automáticamente realizada por el sistema usando unos algoritmos que acomodan el renglón para una producción tipográfica de alta calidad. Estos algoritmos no se encuentran en los procesadores de texto usuales. Además de producciones con contenido matemático, TeX y sus extensiones son aptos para producciones en distintos lenguajes alfabéticos, en música, en química y presentaciones gráficas en general. El costo consiste en el aprendizaje de los comandos del lenguaje pues la mayoría de los programas y editores de texto relacionados son hasta la fecha de dominio público.

Las principales implementaciones de TeX y LaTeX son de código abierto y libre acceso: TeXLive (para sistemas basados en UNIX y Windows), MikTeX

---

[1] Eitan Gurari, fallecido inesperadamente el 22 de junio de 2009, fue el creador, entre otros proyectos, del sistema TeX4ht, ampliamente usado para la conversión de archivos TeX/LaTeX a HTML. Ver obituario en http://root.tug.org/TUGboat/Articles/tb30-2/tb95gurari.pdf.

(Windows), etc. Existen distintas máquinas generadoras que realizan tareas específicas como pdfTeX (esta genera un .pdf listo para impresión definitiva a partir de un archivo .tex), e-TeX, Omega y otras. Hay convertidores de archivos madre .tex en archivos .html para exhibición en la Red. Una información detallada de los recursos de TeX disponibles se encuentra en la página TeX Resources on the Web.

Los Programas gratuitos o comerciales basados en el sistema LaTeX son numerosos. En general estos realizan una compilación de un archivo de texto simple (por ejemplo nota.tex) convirtiéndolo en un archivo final que puede ser un archivo en formato "Postscript": nota.ps, en "Device Independent Format": nota.dvi o en "Portable Document Format": nota.pdf. Algunos de ellos lo hacen al estilo WYSIWYM, esto es, similar a la mayoría de los editores de texto usuales pero introduciendo caracteres propios del programa (que pueden ser texto fuente de TeX, LaTeX o paletas deplegables) uno obtiene un archivo .dvi, .ps o .pdf en formato final: por ejemplo MathType, Scientific Word, Scientific Author, Textures, LyX, Gnu TeXmacs. En la mayoría de los programas gratuitos el estilo de edición es a nivel de código de TeX o LaTeX en el archivo fuente (un archivo de texto con extensión .tex). No obstante con implementaciones de LaTeX capaces de usar "flashmode" se puede conseguir el estilo WYSIWYM en algunos programas gratuitos como TeXShop. La página http://en.wikipedia.org/wiki/Comparison_of_TeX_editors presenta una comparación de los principales programas editores de lenguaje TeX o LaTeX. Un compendio del sistema LaTeX está condensado en el libro The LaTeX Companion de Frank Mittelbach, Michel Goossens et al. Las posibilidades de manejo de gráficos en LaTeX están descriptas en The LaTeX Graphics Companion de Michel Goossens, Sebastian Rahtz y Frank Mittelbach et al. y la conversión de documentos LaTeX a HTML se trata en el siguiente libro: The LaTeX Web Companion de Michel Goossens, Sebastian Rahtz et al.

2. TeX4ht

Se dispone de varias herramientas para la conversión de un archivo .tex a un archivo .html. Una de la más versátiles es TeX4ht que actualmente es mantenida en el sitio http://tug.org/tex4ht/ por CV Radhakrishnan y Karl Berry. Este programa está incorporado en las principales implementaciones de TeX como TeXLive y MikTeX. Una lista bastante extensiva de los conversores disponibles de TeX/LaTeX a HTML/ XML se encuentra en la página http://tug.org/applications/tex4ht/ . Para el uso de TeX4ht se aplican lineas de comando sencillas en una terminal:

```
# htlatex nota1.tex ``html, pic-m´´
```

Con este comando le pedimos a TeX4ht que convierta el archivo nota1.tex a nota1.html y también que todas las fórmulas sean convertidas en imágenes. El programa produce también una hoja de estilo nota1.css y otros archivos auxiliares. El archivo nota1.tex no es un archivo con todas las lineas de comando propias de TeX/ LaTeX. Se pueden introducir algunas líneas de código HTML en forma especial y además el archivo formateado con LaTeX nota1.pdf no luce como un archivo final para imprenta. Es decir, un archivo .tex debe ser retocado antes de ser procesado por TeX4ht y al menos contener en el preámbulo el comando

```
\usepackage[html]{tex4ht}
```

Esto permite además ver si hay algún problema de compatibilidad con los distintos macros usados en el archivo .tex ya que TeX4ht es muy sensible a ciertos comandos usuales en TeX y no es necesariamente compatible con todos los macros y paquetes de TeXLive o MikTeX.

La elección de convertir todas las fórmulas en imágenes se debe a que ellas no se traducen bien al lenguaje HTML. LaTeX es un lenguaje de marcado que ofrece detalles extremos de la disposición de la página mientras que los formatos HTML/XML no. Esto convierte el trabajo de los archivos resultantes para la metodología de SciELO en algo engorroso ya que de un solo trabajo con contenido matemático pueden aparecer cientos de imágenes. Uno podría preguntarse que hay al respecto de los restantes conversores de TeX/LaTeX a HTML/XML, sin embargo un estudio comparativo realizado en el trabajo de Heinrich Stamerjohanns et al. http://kwarc.info/kohlhase/submit/dml09.pdf (ver también la presentación http://www.fi.muni.cz/~sojka/dml-2009-kohlhase.pdf) muestra que entre los principales conversores uno de los más confiables es TeX4ht.

3. MathML

MATHML es un lenguaje de marcado extensión de XML que apunta a integrar fórmulas matemáticas en la Red de Redes WWW. Como una aplicación asociada a XML se pueden considerar secciones, listas, tablas, imágenes, etc. En MathML la presentación y el contenido están separados y puede ser decidido por el usuario. Más aún, el contenido puede ser visto por personas con discapacidades visuales con una conveniente resolución de la pantalla. Para ello es necesario hasta hoy en día introducir plugins (extensiones) a los browsers usuales. Por ejemplo, para Internet Explorer es necesario el plugin MathPlayer, en Firefox hay que instalar la extensión Fire Vox y fuentes adicionales. Ultimamente la mayoría de los browsers poseen algún soporte de MathML a pesar de que este lenguaje está todavía en su adolescencia.

Un documento LaTeX puede convertirse en MathML mediante software de código abierto. Por ejemplo TeX4ht, TTH, Hermes, Tralics, ORCCA (este último es un servicio on-line) tienen la posibilidad de conversión a MathML. Aplicaciones comerciales como MathType o Scientific Word tienen la opción de guardar el archivo en formato MathML. Procesadores de texto de código abierto como Open Office también y programas CAS (Computer Algebra System) comerciales como Mathematica, Maple, Matlab, Mathcad pueden guardar un documento como MathML. Una lista más completa de conversores se encuentra en W3CMathML. Ver también la página principal W3C MathML software. El trabajo arriba mencionado de H. Stamerjohanns et al. contiene una tabla comparativa de conversores a MathML y también se puede apreciar que TeX4ht sobresale con respecto a los demás. Un estudio bastante completo de lo necesario para implementar TeX/LaTeX y TeX4ht en la conversión de documentos con contenido matemático a HTML/XML/MathML se encuentra en el instructivo de Jacek Polewczak http://www.csun.edu/~hcmth008/mathml/acc_tutorial.pdf, o la presentación del mismo http://www.csun.edu/~hcmth008/mathml/converting_to_mathml.pdf .

Los comandos más usuales para la conversión de un archivo .tex a mathml están descriptos en los instructivos de Polewczak o en el de Eitan Gurari HTML

Production . En una instalación de TeXLive para obtener un archivo .html con detección de mathml se puede usar la línea de comando:

htlatex nota1.tex ``html, pmathml´´

No todas las fórmulas son convertidas en texto pero una gran cantidad es detectada por un posprocesador y traducida.

La instalación completa de TeX4ht requiere de compiladores en C y no es tan simple. No obstante permite usar otros comandos como mzlatex, oolatex, xhlatex, dbmlatex que no se encuentran en una instalación usual y con posibilidades interesantes como la conversión en un documento LaTeX a Word, distintas posibilidades de conversión a formatos HTML, XML y MATHML adaptables a las configuraciones particulares de cada browser. Estos comandos están descriptos en los instructivos de Polewczak y Gurari para cada plataforma.

La conveniencia de usar MathML es la de evitar una cantidad desmesurada de imágenes. Las imágenes se obtienen esencialmente de fórmulas resaltadas (entre símbolos $$...$$ en código de TeX) o ecuaciones. La desventaja es que MathML todavía no está completamente implementado en los browsers más usados y que es todavía un emprendimiento en desarrollo. En algunas plataformas como Macintosh se puede también realizar la conversión en forma automática mediante programas de interface gráfica de usuario (GUI) como SimpleTeX4ht.

4. Conclusiones

MathML tiene cierta ventajas respecto de las transformaciones usuales a HTML. Esta siendo adoptado en la generación de archivos de texto completo .html por empresas como Hindawi Publising Coorporation. Desde el punto de vista de documentos LaTeX con contenido matemático se reduce el número de imágenes necesarias y permite catalogaciones con contenido matemático de documentos en red. Usualmente la conversión involucrada requiere de archivos adicionales como hojas de estilo CSS que están hoy en día ampliamente difundidas para documentos HTML. Sin embargo se requiere un gran esfuerzo en sentido retrogrado para llevar estos documentos al formato de SciELO que posee muchas restricciones. Sería deseable que en el futuro el emprendimiento de Bireme/SciELO considerase la ampliación de su sistema a otro más actualizado que en particular considerase la incorporación del lenguaje XML y MathML, así como también archivos adicionales (hojas de estilo, etc.) que son de uso cada vez más frecuente en páginas .html de la Red.

Bahía Blanca, Argentina, 10 de agosto de 2010