Introducción a los Lenguajes de Marca

Orígenes

- Tradicionalmente, en la época de la imprenta, los manuscritos de autor incluían instrucciones que indicaban el tipo de letra, el estilo y el tamaño con que debía ser representado el texto, etc…
- A estas indicaciones se les llamaba marcas, y existía un buen número de ellas conocidas y manejadas informalmente por los tipógrafos

Definición

Un lenguaje de marcado o l**enguaje de marcas** es una forma de codificar un documento que, junto con el texto, incorpora etiquetas o marcas que contienen información adicional acerca de la estructura del texto o su presentación.

Era informática

- Con la introducción de las computadoras, y sobre todo de la web, se trasladó este concepto al mundo de la informática.
- Los ordenadores representan la información mediante un sistema binario (0 y 1). Para ello, previamente ha de ser codificada.
-

Representación de la información

- Datos binarios
  - Cualquier dato que no sea texto, se considera dato binario. Por ejemplo: música, vídeo, imagen, un archivo Excel, un programa, etc.
- Texto plano
  - El texto es quizá la forma más humana de representar información. Esa forma de transmitir es milenaria y sigue siendo la forma más habitual de transmitir información entre humanos; incluso con la tecnología actual aplicaciones como Twitter, Whatsapp, etc; siguen usando el texto como formato fundamental para transmitir información.

Codificación en texto plano

- La forma habitual ha sido codificar cada carácter en una serie de números binarios.
  - el carácter A -> 01000001 y la B el 01000010.
- Estándares para intentar que todo el hardware y software codifique los caracteres igual.
- Conjuntos de caracteres más utilizados:
  - ASCII.
  - ISO 8859.
  - UNICODE.

Ventajas de los archivos binarios

- Ocupan menos espacio. Optimizan
  - el número 213 ocupa un solo byte y no tres
- Son más rápidos de manipular por parte del ordenador.
- Permiten el acceso directo a los datos. Los archivos de texto siempre se manejan de forma secuencial, más lenta.
- En cierto modo permiten cifrar el contenido que de otra forma sería totalmente visible por cualquier aplicación capaz de entender textos (como el bloc de notas).

Ventajas de los archivos de texto

- Son ideales para almacenar datos para exportar e importar información a cualquier dispositivo electrónico ya que cualquier es capaz de interpretar texto.
- Son directamente modificables, sin tener que acudir a software específico.
- Su manipulación es más sencilla que la de los archivos binarios.
- Son directamente transportables y entendibles por todo tipo de redes.

Aparición

- Se ha intentado que los archivos de texto plano (archivos que sólo contienen texto y no otros datos binarios) pudieran servir para almacenar otros datos como, por ejemplo, detalles sobre el formato del propio texto u otras indicaciones.

Procesadores de texto

- Necesitan guardar datos referidos al formato del texto, tamaño de la página, márgenes, etc.
  - Guardar la información de formato de forma **binaria**, lo que provoca los ya comentados problemas.
  - Guardar toda la información como **texto**, haciendo que las indicaciones de formato no se almacenen de forma binaria sino textual. Dichas indicaciones son caracteres marcados de manera especial para que así un programa adecuado pueda traducir dichos caracteres no como texto sino como operaciones que finalmente producirán mostrar el texto del documento de forma adecuada.

Procesadores de texto

- La idea del marcado procede del inglés **marking up** término con el que se referían a la técnica de marcar manuscritos con lápiz de color para hacer anotaciones.
- Las posibles anotaciones o indicaciones incluidos en los documentos de texto han dado lugar a lenguajes (entendiendo que en realidad son formatos de documento y no lenguajes en el sentido de los lenguajes de programación de aplicaciones) llamados lenguajes de marcas, lenguajes de marcado o lenguajes de etiquetas.

Historia de los lenguajes de marcado

- Goldfarb
  - Se considera a Charles Goldfarb como al padre de los lenguajes de marcas. Se trata de un investigador de IBM que propuso ideas para que los documentos de texto tuvieran la posibilidad de indicar el formato del mismo.
  - Al final ayudó a realizar el lenguaje **GML** de IBM el cual puso los cimientos del futuro SGML ideado por el propio Goldfarb.

Historia de los lenguajes de marcado

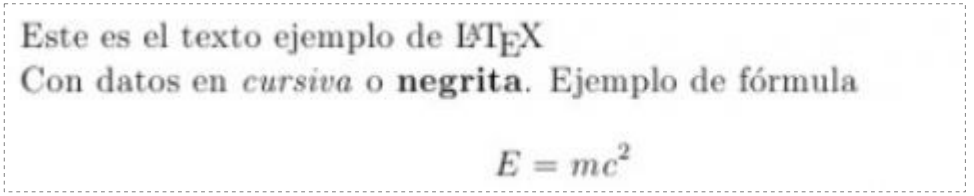
- TeX
  - Década de los 70, Donald Knuth lo creó para producir documentos científicos utilizando una tipografía y capacidades que fueran iguales en cualquier computadora, asegurando además una gran calidad en los resultados.
  - Para ello apoyó a TeX con tipografía especial (fuentes Modern Computer) y un lenguaje de definición de tipos (METAFONT). Ha tenido cierto éxito en la comunidad científica.
  - Crear documentos con tipos de gran calidad, para ello se necesita un programa capaz de convertir el archivo TeX a un formato de impresión.

Historia de los lenguajes de marcado

- LaTeX
  - El éxito de TeX produjo numerosos derivados de los cuales el más popular es (LaTeX). Se trata de un lenguaje que intenta simplificar a TeX, fue definido en 1984 por Leslie Lamport, aunque después ha sido numerosas veces revisado. Al utilizar comandos de TeX y toda su estructura tipográfica, adquirió rápidamente notoriedad y sigue siendo utilizado para producir documentos con expresiones científicas, de gran calidad. La idea es que los científicos se centren en el contenido y no en la presentación.

Ejemplo LaTex

```
\documentclass[12pt]{article}
\usepackage{amsmath}
\title{\Ejemplo}
\begin{document}
Este es el texto ejemplo de \LaTeX{}
Con datos en \emph{cursiva} o \textbf{negrita}.
Ejemplo de f\'ormula
\begin{align}
E &= mc^2
\end{align}
\end{document}
```



Historia de los lenguajes de marcado

- RTF
  - RTF es el acrónimo de Rich Text Format (Formato de Texto Enriquecido) un lenguaje ideado por Microsoft en 1987 para producir documentos de texto que incluyan anotaciones de formato.
  - Actualmente se trata de un formato aceptado como texto con formato y en ambiente Windows es muy utilizado como formato de intercambio entre distintos procesadores por su potencia. El procesador de texto Word Pad incorporado por Windows lo utiliza como formato nativo.

Historia de los lenguajes de marcado

- SGML
  - Se trata de la versión de GML que estandarizaba el lenguaje de marcado y que fue definida finalmente por ISO como estándar mundial en documentos de texto con etiquetas de marcado.
  - La estandarización la hace el subcomité SC24 que forma parte del comité JTC1 del organismo IEC de ISO que se encarga de los estándares electrónicos e informáticos (en definitiva se trata de una norma ISO/IEC JTC1/SC24, concretamente la 8879).

Historia de los lenguajes de marcado

- SGML
  - Es el padre del lenguaje XML y la base sobre la que se sostiene el lenguaje HTML.
  - En SGML las etiquetas que contienen indicaciones para el texto se colocan entre símbolos < y >. Las etiquetas se cierran con el signo /. Es decir las reglas fundamentales de los lenguajes de etiquetas actuales ya las había definido SGML.

Historia de los lenguajes de marcado

- SGML
  - En realidad (como XML) no es un lenguaje con unas etiquetas concretas, sino que se trata de un lenguaje que sirve para definir lenguajes de etiquetas; o más exactamente es un lenguaje de marcado que sirve para definir formatos de documentos de texto con marcas.

Historia de los lenguajes de marcado

- PostScript
  - Se trata de un lenguaje de descripción de páginas. De hecho es el más popular. Permite crear documentos en los que se dan indicaciones potentísimas sobre como mostrar información en el dispositivo final.
  - Es en realidad todo un lenguaje de programación que indica la forma en que se debe mostrar la información que puede incluir texto y el tipo de letra del mismo, píxeles individuales y formas vectoriales (líneas, curvas). Sus posibilidades son muy amplias.

Historia de los lenguajes de marcado

- HTML
  - Tim Bernes Lee utilizó SGML para definir un nuevo lenguaje de etiquetas que llamó ***Hypertext Markup Language*** (lenguaje de marcado de hipertexto) para crear documentos transportables a través de Internet en los que fuera posible el hipertexto; es decir, la posibilidad que determinadas palabras marcadas de forma especial permitieran abrir un documento relacionado con ellas.

Historia de los lenguajes de marcado

- HTML
  - Las páginas web se hacen en HTML.
  - Inicialmente estos documentos se veían con ayuda de intérpretes de texto (como por ejemplo el Lynx de Unix) que simplemente coloreaban el texto y remarcaban el hipertexto. Después el software se mejoró y aparecieron navegadores con capacidad más gráfica para mostrar formatos más avanzados y visuales.

Historia de los lenguajes de marcado

- XML
  - Se trata de un subconjunto de SGML ideado para mejorar el propio SGML y con él definir lenguajes de marcado con sintaxis más estricta, pero más entendibles.
  - Su popularidad le ha convertido en el lenguaje de marcado más importante de la actualidad y en el formato de documentos para exportación e importación más exitoso.

Historia de los lenguajes de marcado

- JSON
  - Abreviatura de JavaScript Object Notation, Se trata de una notación de datos procedente del lenguaje JavaScript estándar (concretamente ECMA Script de 1999).
  - En el año 2002 se le daba soporte desde muchos de los navegadores y su fama ha sido tal que ahora se ha convertido en una notación independiente de JavaScript que compite claramente con XML.

Historia de los lenguajes de marcado

- JSON
  - Se trata de una notación que realmente no se considera lenguaje de marcas, ya que no hay diferencia en el texto a través de etiquetas, sino que se basa en que el texto se divide en dato y metadato. De modo que el símbolo de los dos puntos separa el metadato del dato. Por otro lado los símbolos de llave y corchete permiten agrupar de manera correcta los datos.

Ejemplo de JSON

```
{
  "nombre": "Jorge",
  "apellido1": "Sánchez",
  "dirección":
  {
     "calle": "C/ Falsa nº 0",
     "localidad": "Palencia",
     "código Postal": "34001",
     "país":"España"
  }
}
```

Clases de Lenguaje de Marcado

Se suele diferenciar entre tres clases de lenguajes de marcado, aunque en la práctica pueden combinarse varias clases en un mismo documento.

- Marcado de presentación
- Marcado de procedimiento
- Marcado descriptivo

Orientado a presentación

- El marcado de presentación es aquel que indica el formato del texto.
- Este tipo de marcado es útil para **maquetar** la presentación de un documento para su lectura, pero resulta insuficiente para el procesamiento automático de la información.
  - En ellos al texto común se añaden palabras encerradas en símbolos especiales que contienen indicaciones de formato que permiten a los traductores de este tipo de documentos generar un documento final en el que el texto aparece con el formato indicado.

Orientado a presentación

- Es el caso de HTML en el que se indica cómo debe presentarse el texto (y no por ejemplo lo que significa el mismo) también se considera así los archivos generados por los procesadores de texto tradicionales en los que al texto del documento se le acompaña de indicaciones de formato (como negrita, cursiva, etc.).

Orientado a presentación

- El marcado de presentación resulta más fácil de elaborar, sobre todo para cantidades pequeñas de información. Sin embargo resulta complicado de mantener o modificar, por lo que su uso se ha ido reduciendo en proyectos grandes en favor de otros tipos de marcado más estructurados.

Orientado a procedimientos

- El marcado de procedimientos está enfocado hacia la presentación del texto, sin embargo, también es visible para el usuario que edita el texto.
- El programa que representa el documento debe interpretar el código en el mismo orden en que aparece.
- El archivo en realidad contiene instrucciones a realizar con el texto

Orientado a procedimientos

- Por ejemplo, para formatear un título, debe haber una serie de directivas inmediatamente antes del texto en cuestión, indicándole al software instrucciones tales como centrar, aumentar el tamaño de la fuente, o cambiar a negrita.
- Inmediatamente después del título deberá haber etiquetas inversas que reviertan estos efectos. En sistemas más avanzados se utilizan macros o pilas que facilitan el trabajo.

Orientado a procedimientos

- Algunos ejemplos de marcado de procedimientos son *nroff, troff, TeX*.
- Este tipo de marcado se ha usado extensivamente en aplicaciones de edición profesional, manipulados por tipógrafos calificados, ya que puede llegar a ser extremadamente complejo.

Orientado a la descripción

- El marcado descriptivo o semántico utiliza etiquetas para describir los fragmentos de texto, pero sin especificar cómo deben ser representados, o en qué orden. Los lenguajes expresamente diseñados para generar marcado descriptivo son el **SGML** y el **XML**.

Orientado a la descripción

- Las etiquetas pueden utilizarse para añadir al contenido cualquier clase de metadatos.
  - Por ejemplo, el estándar Atom, un lenguaje de sindicación, proporciona un método para marcar la hora "actualizada", que es el dato facilitado por el editor de cuándo ha sido modificada por última vez cierta información.
  - El estándar no especifica cómo se debe representar, o siquiera si se debe representar. El software puede emplear este dato de múltiples maneras, incluyendo algunas no previstas por los diseñadores del estándar.

Orientado a la descripción

- Una de las virtudes del marcado descriptivo es su flexibilidad:
  - los fragmentos de texto se etiquetan tal como son, y no tal como deben aparecer.
- Estos fragmentos pueden utilizarse para más usos de los previstos inicialmente.
  - Por ejemplo, los hiperenlaces fueron diseñados en un principio para que un usuario que lee el texto los pulse. Sin embargo, los buscadores los emplean para localizar nuevas páginas con información relacionada, o para evaluar la popularidad de determinado sitio web.

Orientado a la descripción

- El marcado descriptivo también simplifica la tarea de reformatear un texto, debido a que la información del formato está separada del propio contenido.
  - Por ejemplo, un fragmento indicado como cursiva (***texto***), puede emplearse para marcar énfasis o bien para señalar palabras en otro idioma. Esta ambigüedad, presente en el marcado presentacional y en el procedimental, no puede soslayarse más que con una tediosa revisión a mano. Sin embargo, si ambos casos se hubieran diferenciado descriptivamente con etiquetas distintas, podrían representarse de manera diferente sin esfuerzo.

Orientado a la descripción

- El marcado descriptivo está evolucionando hacia el marcado genérico.
- Los nuevos sistemas de marcado descriptivo estructuran los documentos en árbol, con la posibilidad de añadir referencias cruzadas. Esto permite tratarlos como bases de datos, en las que el propio almacenamiento tiene en cuenta la estructura, no como en los grandes objetos binarios (blobs) como en el pasado.
- Estos sistemas no tienen un esquema estricto como las bases relacionales, por lo que a menudo se las considera bases semiestructuradas.

Características de los lenguajes de marcas

- Texto plano
  - Compuesto únicamente por caracteres de texto.
- Independencia del dispositivo final (S.O./programa)
- Especialización
  - Se usan en gran variedad de áreas.
- Compacidad
  - Las instrucciones de marcado se mezclan con el propio contenido.
- Flexibilidad
  - Se puede combinar en el mismo archivo con otros lenguajes.