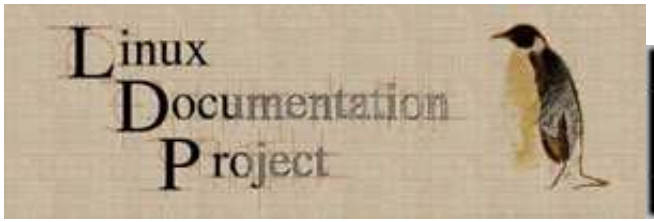
SGML/XML para autores de documentación

Por Ismael Olea <ismael@olea.org>

Una conferencia CACLE



Objetivos

- Comprender las ventajas de la documentación estructurada y la marcación semántica.
- Conocer el entorno de producción de documentación en software libre.

Los documentos tienen estructura

- El documento es la forma generalizada de intercambio de información escrita entre los seres humanos.
- Los documentos se crean mediante convenios lingüísticos:
  - Gramáticos
  - Semánticos
  - Ortográficos
  - Ortotipográficos

    Ejemplo:

    el caso de las comillas españolas «»
- Los documentos contienen información. Como tal, la información es pura, abstracta.
- La información de los documentos está estructurada según unos convenios -ortotipografía- que pueden ser explícitos o implícitos.
- Para que el proceso de comunicación tenga lugar hay que asegurar la máxima complitud de los convenios.

Ventajas de la documentación estructurada

Formato ≠ Estructura

Múltiples formatos de salida de un mismo original.

Marcación descriptiva versus marcación procedimental

La estructuración añade información semántica sobre qué es cada parte en lugar de indicar cómo representarlo.

Sencillez de la creación

El autor se centra en la labor de escritura y creación, sin distraerse con asuntos que no son relevantes en el proceso de autoría.

Independencia y portabilidad

El documento es independiente de la plataforma, facilitando su intercambio, almacenaje, consulta y proceso.

Sencillez de edición

Al tratarse de puro texto es de fácil manejo con cualquier clase de herramientas básicas de edición de textos. Al ser absolutamente estructurado puede usarse cualquier editor estructurado compatible.

Mejora de calidad en los formatos de salida/reproducción

Apoyado por un adecuado sistema de representación mejora la calidad final de los documentos y unifica el uso de las convenciones de estilo de cada idioma y área de conocimiento.

Procesable automáticamente: compatible con la Web Semántica

Gracias al riguroso diseño también se facilita la automatización de procesamiento, la alimentación de aplicaciones de conocimento y el intercambio de documentos por medios informáticos.

Web Semántica

Definición: La **Web Semántica** es la representación abstracta de los **datos** en la World Wide Web usando los estándares XML, RDF y otros por definir. Está siendo desarrollada por el W3C en colaboración con un gran número de investigadores y socios industriales.

"La Web Semántica es una extensión de la web actual en la que el significado de la información queda bien-definido y ésta se optimiza para ser manejada cooperativamente tanto por ordenadores como por personas." -- *Tim Berners-Lee, James Hendler, Ora Lassila, The Semantic Web, Scientific American, May 2001*

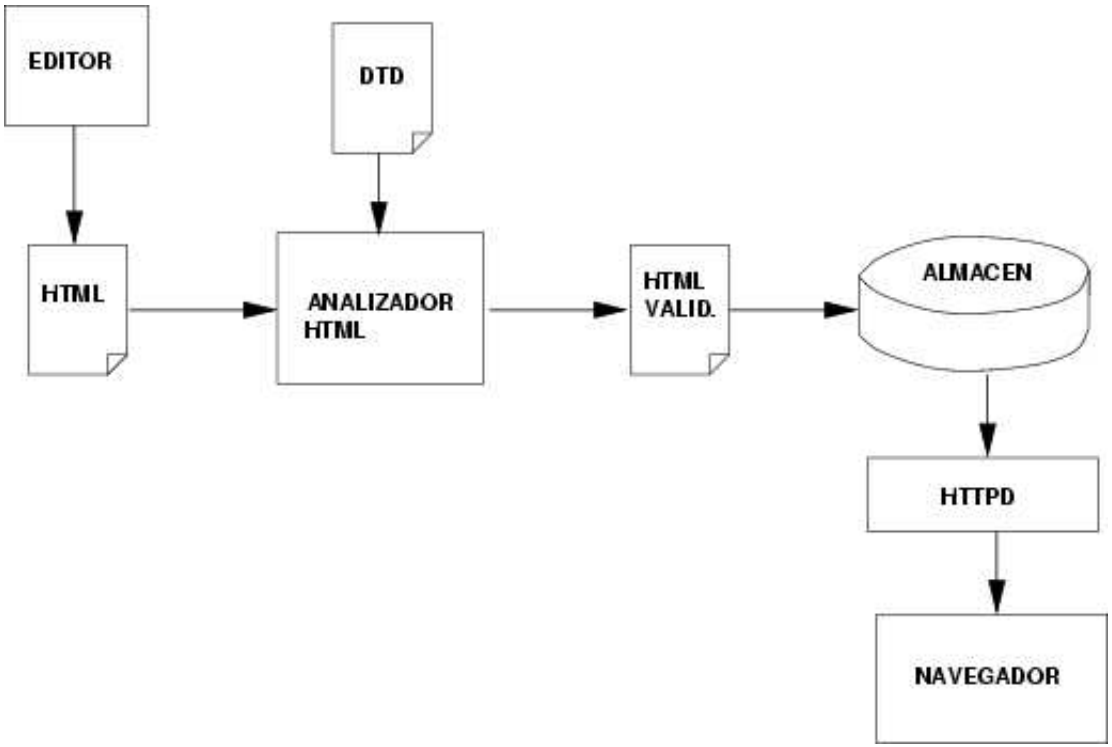
Todos usamos SGML/XML

HTML (HyperText Markup Language) es una tecnología para estructurar documentos genéricos:

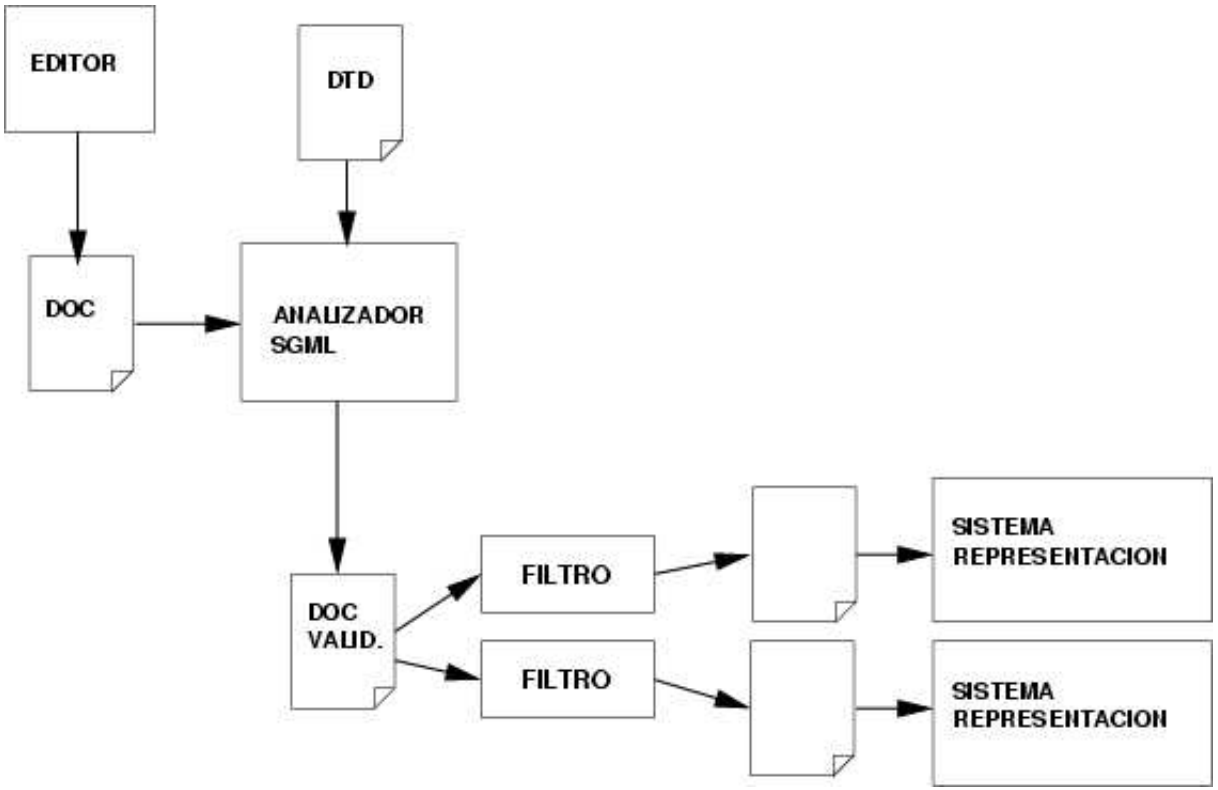
- HTML, expresado en SGML (http://www.w3.org/TR/html401)
- XHTML, expresado en XML (http://www.w3.org/TR/xhtml1/)

- Un ejemplo de HTML
- Un ejemplo de HTML (código fuente)

Entorno de desarrollo HTML



Entorno de producción SGML clásico



SGML

- Standard Generalized Mark-up Language
- Estándar internacional: ISO 8879/1986
- Implementa las ventajas propias de la estructuración de documentos:
  - Independiente de la representación
  - Independiente de la plataforma
  - Abstracción pura de la información
- Creado por el programa CALS del DoD su uso está generalizado en la industria aeroespacial y de defensa.
- Madurado desde 1960 (aplicación de composición de textos GML de IBM).
- Charles Goldfard, su inventor, creó la filosofía de los lenguajes de marcas:

  ...no limita los documentos a una única aplicación, a un estilo de formateado o a un sistema de procesamiento. Se basa en dos postulados novedosos (en aquel momento):

1. *El etiquetado debería describir la estructura del documento y otros atributos más que especificar el procesamiento que se va a llevar a cabo en dicho documento ya que el etiquetado descriptivo necesita efectuarse tan sólo una vez, siendo ésta suficiente para todos los procesamientos futuros.*
2. *El etiquetado debería ser riguroso de manera que las técnicas disponibles para el procesamiento de objetos rigurosamente definidos, como por ejemplo los programas y bases de datos, puedan utilizarse también para el procesamiento de documentos.*

Estructura de SGML

Juego de caracteres

ASCII < Latin1 (ISO 8859-1) < Unicode (ISO/IEC 10646)

El juego de caracteres puede declararse explícitamente o, en su omisión, usar el juego predeterminado por el sistema.

Declaración de documento

La DTD (declaración de tipo de documento) especifica la sintaxis y la jerarquía y relación entre las marcas en las diferentes formas en las que puede construirse un tipo o familia concreto de documentos.

Como símil puede pensarse que es la definición rigurosa de un formulario especialmente flexible.

Elementos

Las secciones que componen al documento desde el punto de vista estructural. La jerarquía y relaciones de los elementos está definida en la DTD.

Familiarmente también se les llama etiquetas o marcas.

Existen dos tipos: «inline» y bloque.

Atributos

Los parámetro de cada elemento. Pueden ser opcionales (y tener o no valores predeterminados) u obligatorios.

Entidades

De uso múltiple y flexible:

- representación de signos no recogidos por el juego de caracteres
- abreviaturas o macros en el código fuente
- referencia a ficheros externos (como los «#includes» del lenguaje C)
- variables, cuyo valor se especifica en el momento del procesamiento.

Las entidades deben estar declaradas en el documento o en la DTD.

Contenido

El texto introducido entre marcas

Instancia

El documento SGML compuesto usando una DTD, sus elementos y los atributos de estos, las entidades internas/externas y el contenido entre marcas.

XML vs. SGML

- Extensible Markup Language.
- Estándar: http://www.w3.org/TR/REC-xml
- XML está escrito en SGML
- XML es funcionalmente equivalente a SGML
- es más sencillo
- no es tan potente:
  - uso obligatorio de comillas en los atributos
  - distingue mayúsculas de minúsculas en los nombres de los elementos
  - no permite restricciones en el anidado de los elementos
  - no permite la minimización de elementos
- Nuevas normas asociadas (algunas rompen compatibilidad SGML):
  - Namespaces
  - XLink
  - XPointer
  - Schemas
  - Xincludes

- RDF
- XSL/XSLT, etc

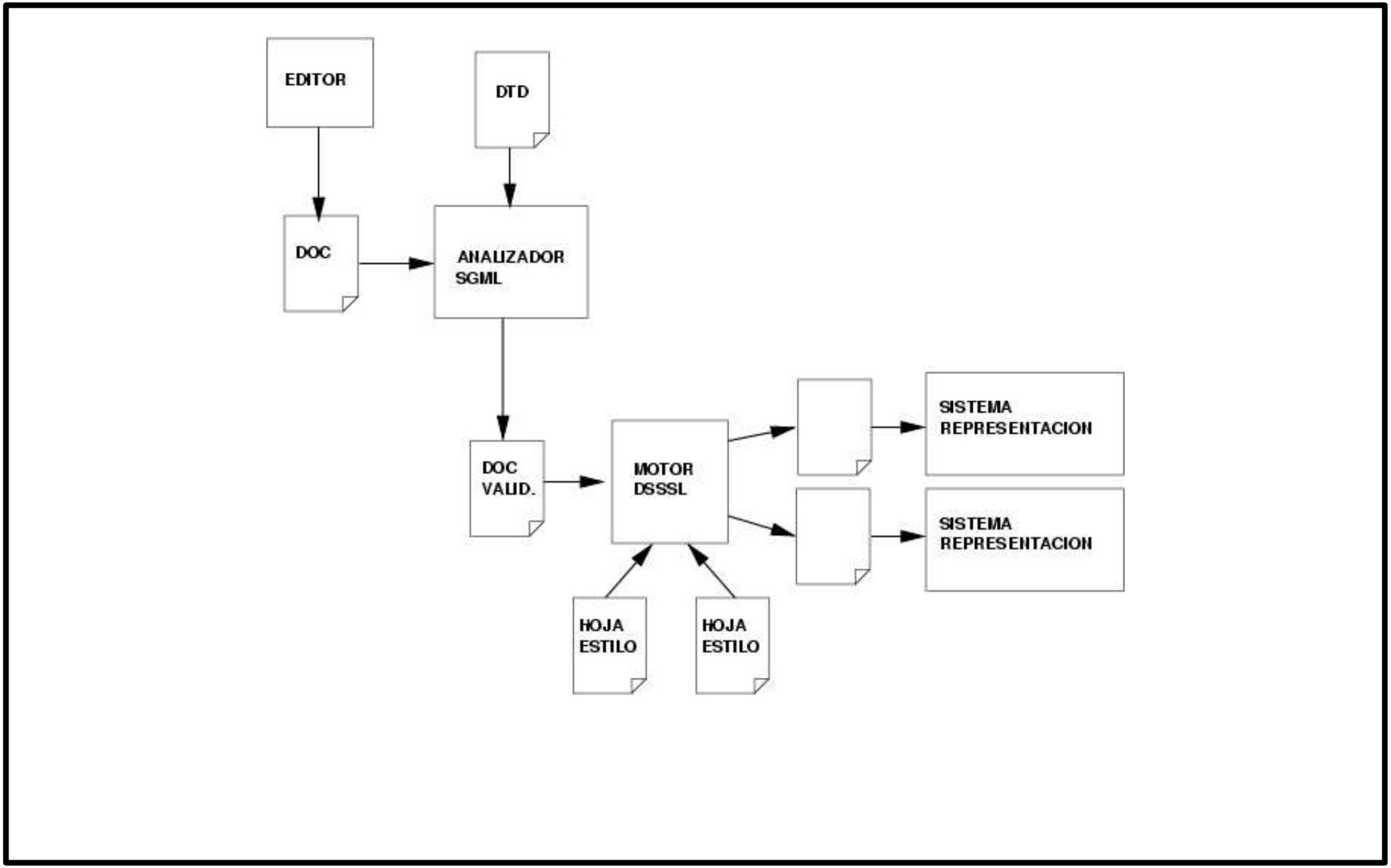
DSSSL

- Document Style Semantics Specificacion Language
- Estándar internacional ISO/IEC 10179:1995
- Salto de gigante en la manipulación de documentos:
  - procesamiento de documentos para su representación;
  - transformación de documentos de una a otra DTD;
  - base para la gestión documental completa.
- Proporciona:
  - entorno uniforme de manejo y modificación de documentos;
  - entorno de creación de filtros mediante hojas de estilo
  - entorno de programación uniforme, basado en Scheme (lenguaje imperativo maduro, robusto, sencillo y perfectamente especificado -R4RS-)

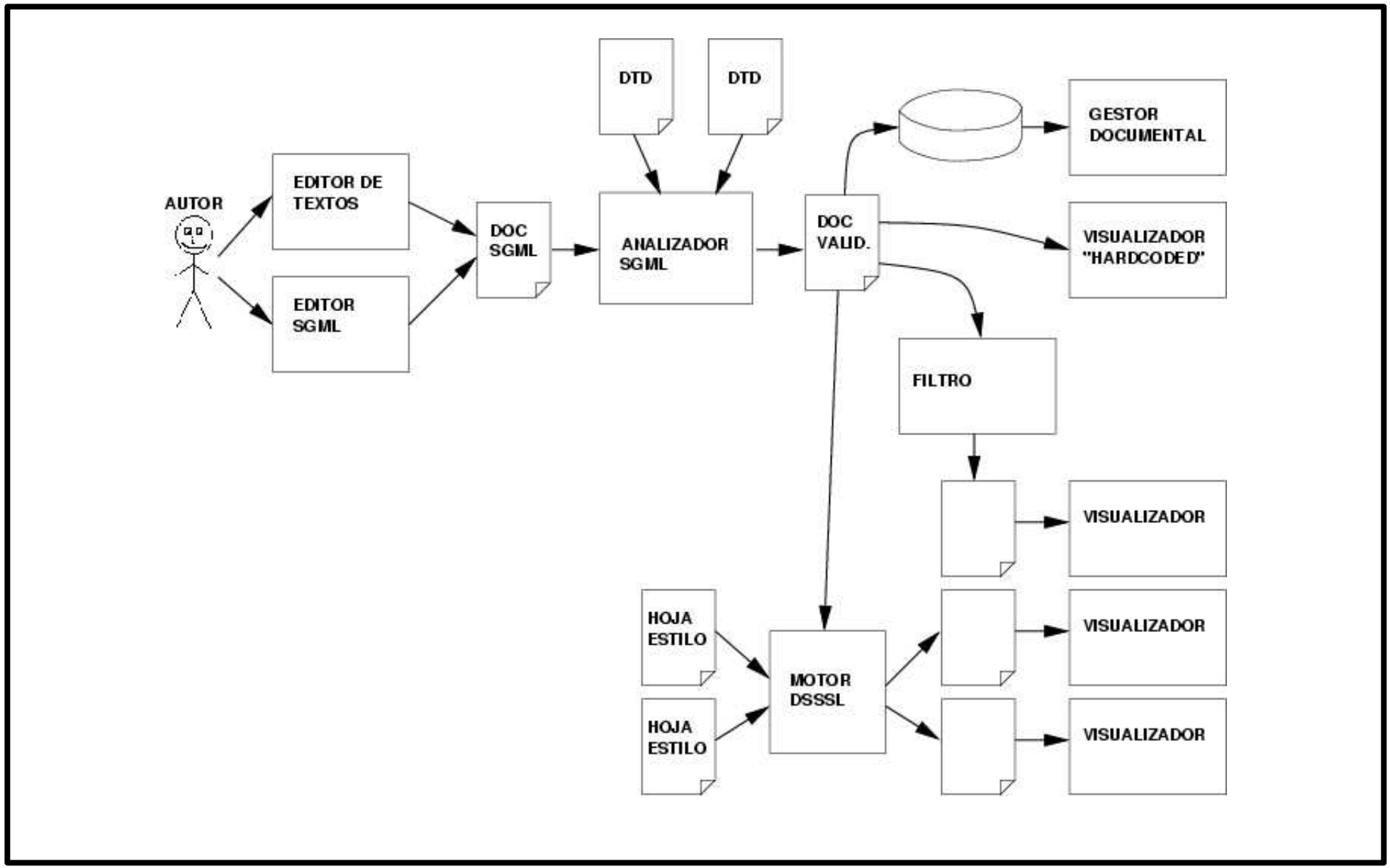
XSL

- Extensible Stylesheet Language
- http://www.w3.org/TR/xsl/
- Basado en JavaScript
- Reimplementación del concepto de hojas de estilo (CSS2 y DSSSL) con el objetivo de más sencillez
  - en la implementación
  - en su uso
- Se compone de:
  - Formateo: procesamiento de documentos para su representación;
  - Trasformación: transformación de documentos de una a otra DTD (**XSLT** http://www.w3.org/TR/xslt);

Entorno de producción SGML ampliado con DSSSL


EDITOR
DTD
DOC
ANALIZADOR SGML
DOC VALID.
MOTOR DSSSL
SISTEMA REPRESENTACION
SISTEMA REPRESENTACION
HOJA ESTILO
HOJA ESTILO

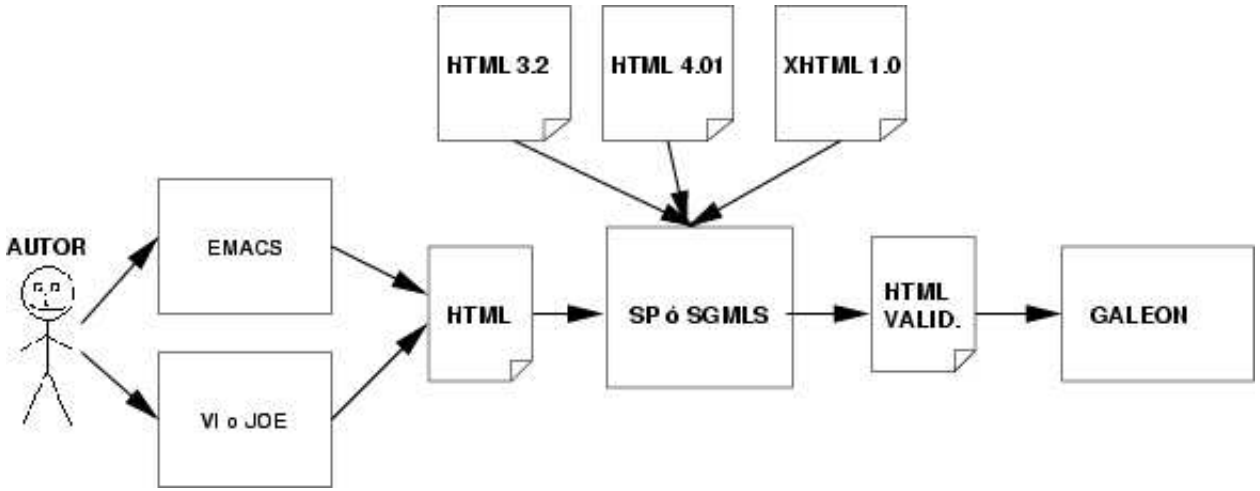
Entorno SGML genérico


DTD
DTD
GESTOR DOCUMENTAL
AUTOR
EDITOR DE TEXTOS
DOC SGML
ANALIZADOR SGML
DOC VALID.
VISUALIZADOR "HARDCODED"
EDITOR SGML
FILTRO
VISUALIZADOR
HOJA ESTILO
MOTOR DSSSL
VISUALIZADOR
HOJA ESTILO
VISUALIZADOR

DTD de éxito: HTML

- **HyperText Mark-up Language**. El sistema SGML más usado de la historia con cientos de millones de usuarios.
- Nació en 1990 para uso específico en la WWW.
- Mantiene carácterísticas de estructuración de documentos a la vez que usa marcas de maquetación.
- La abundancia de navegadores y herramientas existentes no sólo no han respetado completamente los estándares sino que los han ampliado por su cuenta.
- La versión 3.2 de la DTD fue publicada en mayo de 1996. La última versión en vigor es la 4.0 desde principios del 98. Ésta incluye mejoras como el uso de Unicode en el juego de carácteres.
- Las DTD de nivel 4.0 están diseñadas para la implantación progresiva de hojas de estilo: inicialmente CSS, luego CSS2 y especialmente XSL/XSLT.
- Disponible en versión XML: XHTML y en otros derivados de esta versión.

Entorno HTML en Linux



DTD de éxito: CALS

- Continuous Acquisition and Lifecycle Support, programa del DoD bajo cuyos auspicios se desarrolló la tecnología SGML. Su objetivo consistía en simplificar al máximo los formatos de la documentación a fin de efectuar intercambios de datos entre el gobierno y los contratistas de la industria de la defensa.
- Es el entorno más desarrollado dentro de la industria y ha servido de revulsivo para el resto del mundo SGML.
- Entre sus diversas DTD, el sistema de declaración de tablas ha trascendido a otras familias SGML, tales como DocBook.

DTD de éxito: TEI

- Text Encoding Initiative
- Ambicioso proyecto surgido del mundo de las humanidades y las ciencias sociales para desarrollar una DTD común para la codificación e intercambio de documentos relevantes: textos en prosa, verso, drama, manuscritos, etc.
- Aun con toda la tremenda complejidad del proyecto ya están disponibles varias DTD. La más simple se conoce como TEIlite.

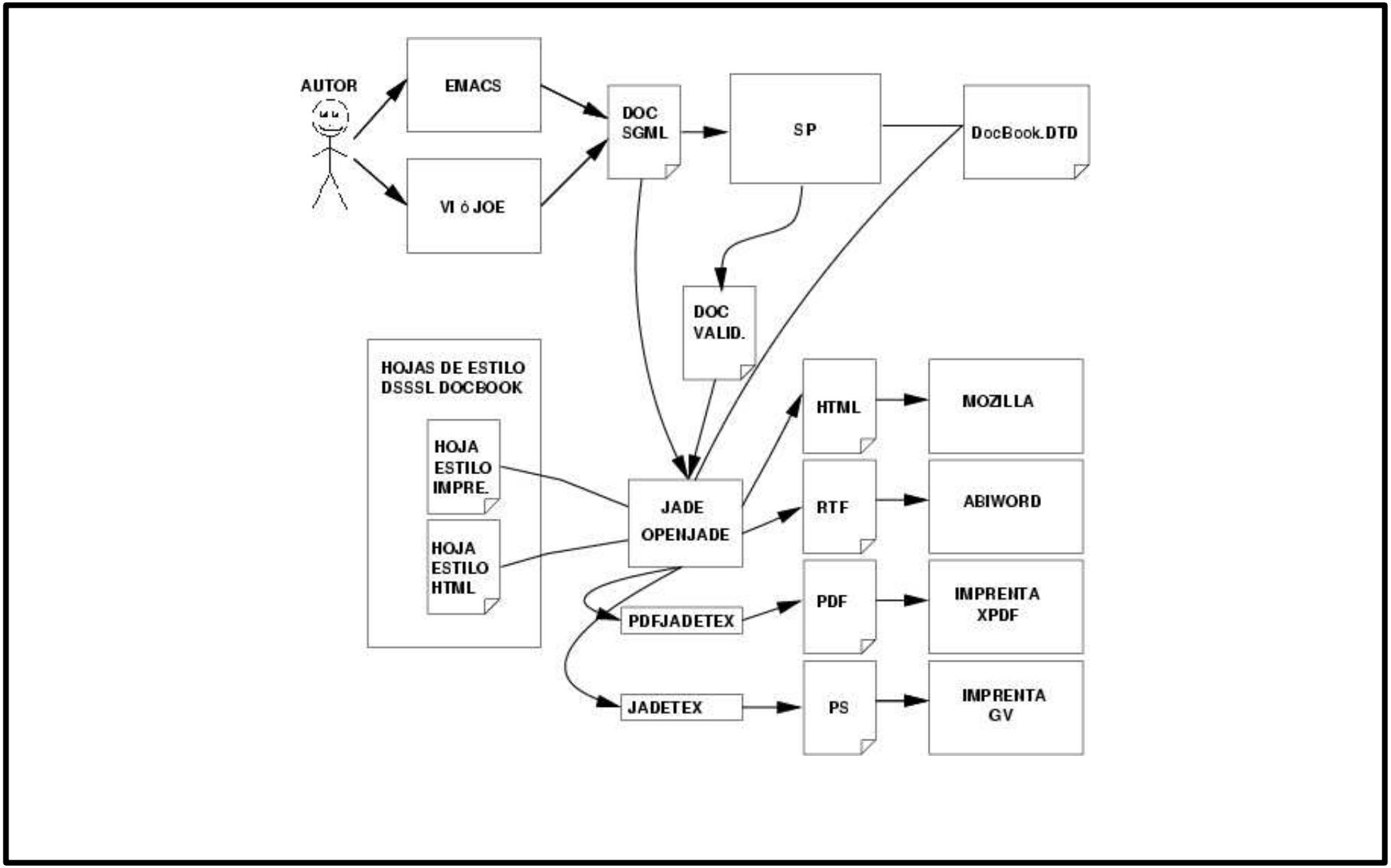
DTD de éxito: Linuxdoc-SGML

- Usado en la estandarización de los Cómos/Howtos.
- Mal llamado «formato SGML».
- Obsoleto por Docbook y en vías de extinción.

DTD de éxito: Docbook

- DTD técnica de informática
- libros, artículos, páginas de manual y colecciones de los mismos
- diseñado por la industria
- actualizado
- SGML/XML
- DSSSL/XSL
- estandarizado en todos los proyectos de soft libre importantes
- incluido en todas las distribuciones modernas

Entorno Docbook SGML en Linux


AUTOR
EMACS
VI ó JOE
DOC SGML
SP
DocBook.DTD
DOC VALID.
HOJAS DE ESTILO DSSSL DOCBOOK
HOJA ESTILO IMPRE.
HOJA ESTILO HTML
JADE OPENJADE
HTML
MOZILLA
RTF
ABIWORD
PDF
IMPRENTA XPDF
PDFJADETEX
JADETEX
PS
IMPRENTA GV

Ejemplo de uso de Docbook

Obtener PDF:

```
docbook2pdf ejemplos/apannao.xml
```

Obtener HTML:

```
docbook2html ejemplos/apannao.xml
```

Herramientas:

jade/**openjade**
: procesador DSSSL (integra el analizador SGML **SP**)

jadetex/**pdfjadetex**
: Interfaz TeX para formatos de reproducción desde **jadetex**

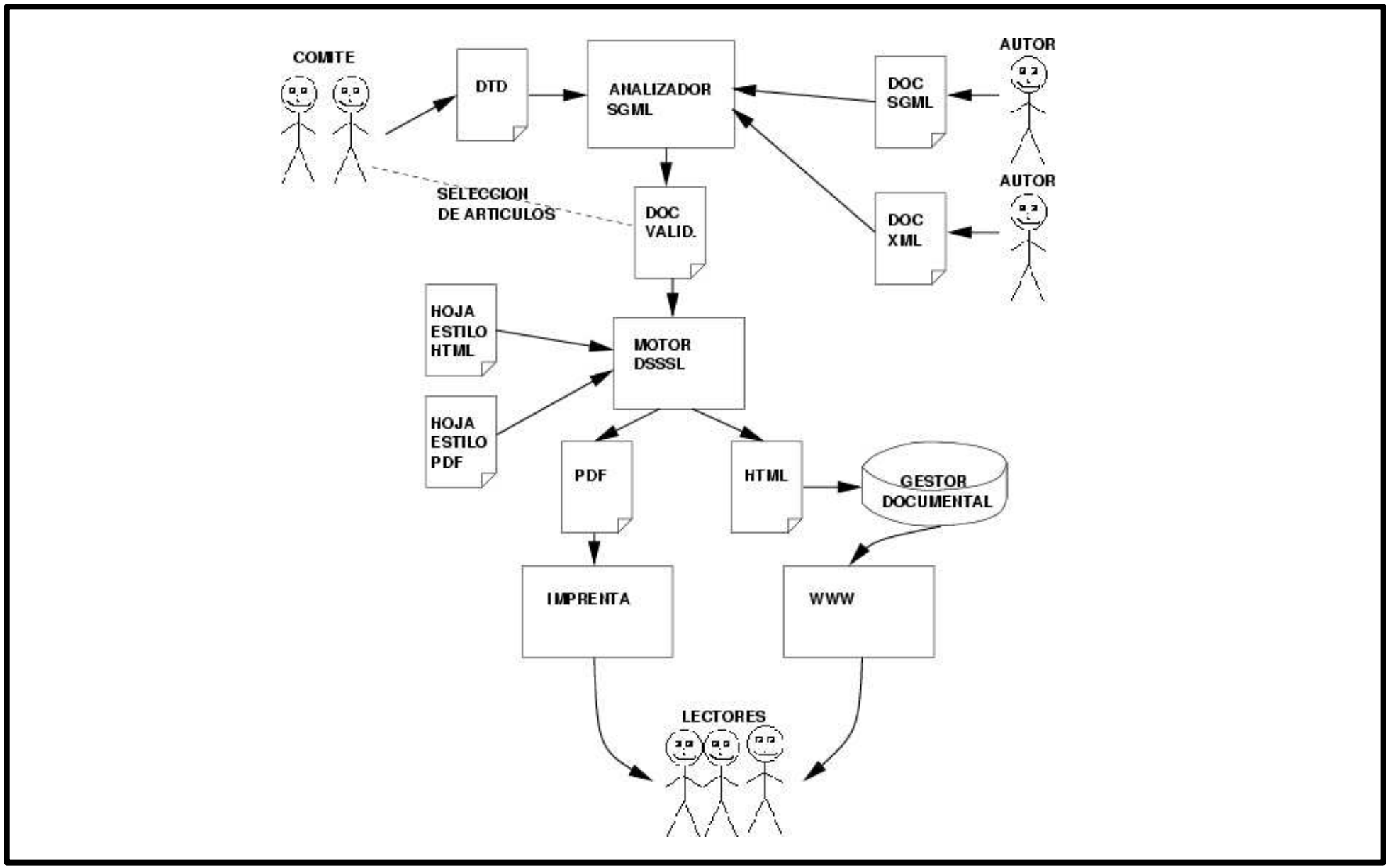
xmllint
: Analizador XML

xsltproc
: Procesador XSLT

Creando documentos estructurados:

1. elección de la DTD
2. planificación del tipo de documento
3. estructuración del documento
4. dar contenido
5. refine del marcado
6. «impresión» final

Ejemplo de uso de aplicación SGML


COMITE
DTD
ANALIZADOR SGML
DOC SGML
AUTOR
AUTOR
SELECCION DE ARTICULOS
DOC VALID.
DOC XML
HOJA ESTILO HTML
HOJA ESTILO PDF
MOTOR DSSSL
PDF
HTML
GESTOR DOCUMENTAL
IMPRENTA
WWW
LECTORES

Preguntas

Hagan alguna pregunta si no quieren descorazonar a este conferenciante

Despedida y cierre

¡Esto es to... esto es to... esto es todo amigos!