En los años 70 continúa la evolución de los lenguajes de programación a la vez que surgen otros lenguajes informáticos orientados a la gestión de información.

Con el desarrollo de los programas que procesan texto surgen los primeros lenguajes informáticos especializados en tareas de descripción y estructuración de información: los lenguajes de marcas.

Uno de los problemas presentes en el desarrollo de este tipo de lenguajes era la falta de estandarización.

Para resolverlo, en los años sesenta IBM encargó a Charles F. Goldfarb, la construcción de un sistema de edición, almacenamiento y búsqueda de documentos legales. Tras analizar la empresa llegaron a la conclusión de que lo mejor era crear un formato estándar para todos los documentos que manejaba la empresa.

El nombre de este formato se llamo GML, cuyo objetivo fue describir los documentos independientemente de la plataforma.

El GML evoluciono hasta 1986 que se creo el estándar ISO 8879 o SGML.

SGML introduce 3 conceptos básicos:

Definir lenguajes (metalenguaje) mediante un DTD (Document Type Definition).

Marcado descriptivo (frente marcado procedural. El marcado descriptivo describe, mediante las marcas o etiquetas definidas en la DTD, la estructura lógica de la información. La idea clave es que las marcas no determinan el procesamiento del documento de manera fija, ya que dicho procesamiento se determina a partir de las necesidades concretas, y se beneficia de la estructura lógica del documento caracterizada a través de sus marcas;)

Independencia de la plataforma. Como los documentos SGML únicamente contienen texto, éstos pueden ser procesados en distintas plataformas, trascendiendo el uso de dichos documentos a los sistemas que los crearon y utilizaron originariamente.

En 1989/90 Tim Berners-Lee se encontró con la necesidad de organizar, enlazar y compatibilizar gran cantidad de información procedente de diversos sistemas. Conociendo SGML creó a partir de su sintaxis un lenguaje de descripción de documentos llamado HTML como combinación de dos estándares.

• ASCII: Es el formato que cualquier procesador de textos sencillo puede reconocer y almacenar. Por tanto, es un formato que permite la trasferencia de datos entre diferentes ordenadores.

• SGML: Lenguaje que permite dar estructura al texto, resaltando los títulos o aplicando diversos formatos al texto.

HTML es una versión simplificada de SGML, ya que sólo se utilizaban las instrucciones absolutamente imprescindibles. Era tan fácil de comprender que rápidamente tuvo gran aceptación. Se crea la World Wide Web y HTML se convierte en el lenguaje para la creación de sus documentos.

En 1998 surge el estándar internacional XML, un lenguaje de marcas puramente estructural que no incluye ninguna información relativa al diseño, permite crear etiquetas adaptadas a las necesidades.

Un lenguaje de marcas es un LENGUAJE que incorpora en el contenido un conjunto de etiquetas o marcas que añaden al documento información adicional acerca de su estructura, presentación, etc... El lenguaje de marcas es el que especifica cuáles serán las etiquetas posibles, donde deben colocarse y el significado que tendrá cada una de ellas. Permiten hacer explícita la estructura de un documento, su contenido semántico o cualquier otra información lingüística que se quiera hacer patente.

No hay que confundir lenguaje de marcas con lenguaje de programación.

Características.

- Se crean con editores de texto, usando archivos de texto plano.
- Permiten la utilización de metadatos.
- Utilizan marcas para incorporar los metadatos al documento.
- Compacidad. Las marcas aparecen junto al contenido del documento.
- Fáciles de interpretar y procesar.
- Flexibilidad.

Clasificación.

Aunque en la práctica, en un mismo documento pueden combinarse varios tipos diferentes de lenguajes de marcas, éstos se pueden clasificar en tres grupos:

Orientados a la presentación: Son los utilizados generalmente por los procesadores de texto y codifican cómo ha de presentarse el documento, es decir, define el formato del texto.

De procedimiento: Orientados también a la presentación, dentro de un marco procedural que permite definir secuencias de acciones, el programa que representa el documento debe interpretar el código en el mismo orden en que aparece.

Descriptivos o semánticos: Este tipo no define qué se debe hacer con un trozo del documento, sino que, las marcas sirven para indicar qué es esa información, es decir, describen qué es lo que se está representando. Describen las diferentes partes, pero sin especificar cómo deben representarse.

XML. Estructura y sintaxis.

XML, siglas de extensible markup language, es un metalenguaje de marcas desarrollado por el World Wide Web Consortium (W3C) , que es un consorcio internacional que crea recomendaciones para la World Wide Web.

XML proporciona una serie de reglas para que cualquiera pueda definir un lenguaje, describiendo su propio conjunto de etiquetas, atributos y relaciones entre estas etiquetas. Al conjunto de reglas de un lenguaje se le llama gramática del lenguaje.

Así, XML es un lenguaje de marcas puramente estructural que no incluye ninguna información relativa al diseño.

Características.

- Permitir definir etiquetas propias.
- Permitir asignar atributos a las etiquetas.
- Utilizar un esquema para definir de forma exacta las etiquetas y los atributos.
- La estructura y el diseño son independientes.

Estructura.

Todo documento XML tiene una estructura jerárquica arborescente y está compuesta por dos partes fundamentales: prólogo, y cuerpo.

El prólogo es la primera parte del documento y contiene la información (meta información) sobre el resto del documento, como versión XML y juego de caracteres utilizado. También puede incluir la referencia a la gramática, recogida en un DTD (Document Type Definition) o en un Schema XML.

El cuerpo del documento se compone de un conjunto de elementos donde cada elemento está formado por:

- **Etiqueta de inicio.** Delimitada por los caracteres < >, incluye nombre de elemento y conjunto de pares atributo, valor.
- **Contenido.** Puede ser texto u otros elementos.
- **Etiqueta final.** Nombre del elemento delimitado por los caracteres </ >

  Para que un documento XML se considere correcto debe estar bien formado y ser válido.

Documentos bien formados: Un documento XML se considera bien formado si cumple las siguientes características o reglas sintácticas:

- Aunque no es obligatorio, W3C recomienda empezar con una declaración XML en la que se indique la versión y el juego de caracteres empleado.
- Estructura jerárquica arborescente. Tiene un único elemento raíz.
- Los elementos deben anidarse correctamente.
- Las etiquetas de cierre de los elementos son obligatorias.
- Se distinguen las mayúsculas y las minúsculas.
- Se permiten elementos vacíos.
- Los valores de los atributos van entrecomillados.
- Los nombres de los atributos deben ser identificadores válidos.
- Los comentarios se deben escribir entre <!-- y -->.

Documentos válidos: Un documento XML se considera válido si está bien formado y además cumple la gramática definida para el lenguaje. La gramática del lenguaje se puede describir por medio de DTD, XML Schema o más recientemente Relax NG.

Espacios de nombres en XML.

Un espacio de nombres XML es una recomendación W3C para proporcionar elementos y atributos con nombre único en un archivo XML. Un archivo XML puede contener nombres de elementos o atributos procedentes de más de un vocabulario XML. Si a cada uno de estos vocabularios se le da un espacio de nombres, un ámbito semántico propio, referenciado a una URL donde se listen los términos que incluye, se resuelve la ambigüedad existente entre elementos o atributos que se llamen igual, la homonimia. Los nombres de elementos dentro de cada espacio de nombres deben ser únicos.