CAPÍTULO I.3

LA DOCUMENTACIÓN JURÍDICA EN LA ERA DE INTERNET

I.3.1. INTRODUCCIÓN A LA RED INTERNET[130]

La aparición de Internet ha supuesto, sin lugar a dudas, una de la mayores revoluciones en el mundo de la informática y de las comunicaciones desde la invención del teléfono, extendiéndose sus implicaciones a todos los órdenes de la vida.

Muchas han sido las definiciones que se han venido aportando sobre Internet desde el surgimiento de esta inmensa red informática, denominada humorísticamente por muchos autores como "la madre de todas las redes", algunas de ellas abordando sus aspectos más técnicos y otras, por el contrario, incidiendo en sus implicaciones sociales, políticas y económicas. Y esto es así, como acertadamente expresan J. Díez y J. De Yraolagoitia, dado que, por un lado, Internet conecta físicamente a ordenadores de todo el mundo y, por el otro, es en la actualidad el medio más común para acceder a un fondo mundial con los recursos y conocimientos de millones de personas e instituciones[131]. La IAB (*Internet Architecture Board*) en uno de sus documentos de trabajo, el RFC 1310, de marzo de 1992, definió a la red Internet como:

> *Un espacio de colaboración internacional vagamente organizado de redes autónomas interconectadas, las cuales soportan comunicaciones host a host a través de una adhesión voluntaria a protocolos y procedimientos abiertos definidos por los estándares de Internet [...] El principal conjunto de estándares de Internet es habitualmente conocido como el TCP/IP protocol suite*[132].

---

[130] Para un mayor y más amplio conocimiento del origen y evolución de la red Internet recomendamos la consulta de los recursos electrónicos referenciados en el sitio web de la ISOC, *All About Internet: Internet Histories*, en la dirección http://www.isoc.org/internet/history/, y, de forma muy especial, la lectura del artículo escrito por un gran número de investigadores que participaron activamente (y aún siguen participando) en la construcción de los fundamentos tecnológicos de Internet. Véase en Barry M. Leiner, Vinton G. Cerf, David D. Clark, Robert E. Kahn, Leonard Kleinrock, Daniel C. Lynch, John Poste, Lawrence G. Roberts, Stephen Wolf. *A Brief History of Internet* [documento HTML]. Internet Society, rev. 2001. Disponible en http://www.isoc.org/internet/history/brief.html (consultado el 17 de enero de 2001). Este artículo fue traducido a nuestro idioma y publicado en formato impreso en *Cuadernos NOVÁTICA*, nº 1, febrero 1999, pp. 1-13.

[131] Jesús Díez y Jaime de Yraolagoitia. "Internet: Red de redes". *PC World*, nº 106, enero 1995, p. 129.

[132] Lyman Chapin. *Request for Comments 1310: The Internet Standards Process* [documento TXT]. Network Working Group, IAB, March 1992. Disponible en ftp://ftp.isi.edu/in-notes/rfc1310.txt (consultado el 13 de noviembre de 2000).

Pero, sin lugar a dudas, la definición más ajustada, exacta y autorizada la proporciona la *Internet Society* (ISOC), al exponer que:

> *La Internet es una red global de redes que permite a toda clase de ordenadores comunicarse y compartir servicios de forma directa y transparente a través de buena parte del mundo. Puesto que Internet es un potencial enormemente valioso y que ofrece tantas posibilidades para tantas personas y organizaciones, también constituye un recurso global y compartido de información y conocimiento, y un medio de colaboración y cooperación entre innumerables comunidades diferentes*[133].

Miguel A. Sanz emplea cuatro adjetivos para definir magistralmente el significado y alcance de la red Internet; éstos son[134]:

- Grande: se trata de la mayor red de ordenadores existente en todo el mundo y la que más deprisa crece.
- Cambiante: se trata de una red informática en continua evolución y adaptación a las nuevas necesidades y circunstancias tecnológicas y sociales.
- Diversa: en ella se integran todo tipo de equipos informáticos, fabricantes de hardware y de software, redes de comunicación, y medios físicos de transmisión, de usuarios, etc.
- Descentralizada: se trata del mejor ejemplo existente de trabajo cooperativo dado que no existe una autoridad central o propietaria de dicha red, sino, más bien, un esfuerzo de colaboración entre los organismos nacionales que controlan cada una de las redes informáticas integradas.

Haciendo una breve revisión histórica de la red Internet, se puede situar su origen a finales de la década de los sesenta con el desarrollo de un proyecto de red informática

---

[133] Traducción tomada de J. Tomás Nogales Flores. "Los usos básicos de Internet: Servicios y aplicaciones". En: Mercedes Caridad Sebastián (coord.). *La Sociedad de la Información: Políticas, Tecnología e Industria de los contenidos*. Madrid: Centro de Estudios Ramón Areces: Universidad Carlos III de Madrid, 1999, p. 143, realizada del documento original *What is Internet?* [documento HTML]. Internet Society, rev. 2001. Disponible en http://www.isoc.org/internet/index.shtml (consultado el 23 de enero de 2001). A estas tres vertientes básicas de la naturaleza de Internet (hardware, software e información), J. T. Nogales añade asimismo una cuarta: el *factor humano* o todas aquellas personas que están al otro lado de estos componentes de la Red, factor determinante para comprender su verdadera naturaleza más allá de lo meramente tecnológico.

[134] Miguel A. Sanz Sacristán. "A, B, C de Internet". *Boletín de RedIRIS*, nº 28, julio 1994, p. 16.

experimental para la DARPA (*Defense Advanced Research Projects Agency*) del Departamento de Defensa de los Estados Unidos. Esta nueva red se basaba en ideas y estudios anteriores de algunos investigadores del prestigioso *Massachusetts Institute of Technology* (MIT) en donde se debatía sobre el concepto y posibilidades de construcción de una red informática de alcance mundial (*Galactic Network*) y de la importancia del concepto de trabajo en red. Varios fueron los objetivos que animaron el desarrollo de esta red experimental y que constituyeron toda una auténtica revolución en la época; entre éstos podemos destacar los tres siguientes:

1. **Tecnología abierta**: construir un sistema de comunicaciones entre ordenadores que fuese lo suficientemente flexible y dinámico para permitir el uso de cualquier tipo de medio y tecnología de transmisión.

2. **Red descentralizada**: basarse en el concepto de red informática distribuida (no existen nodos centrales). Este hecho era de importancia vital en la época dado que se estaba en el periodo conocido como "guerra fría" y, por tanto, esta red debería ser inmune a eventuales bombardeos nucleares de tal modo que aunque cayese un nodo de dicha red, el tráfico de datos no se viese interrumpido, buscando otro modo o ruta alternativa para llegar a su destino[135].

3. **Conmutación de paquetes**: debería basarse en la conmutación de paquetes frente a las redes imperantes en la época de conmutación de circuitos. La idea era que para aumentar la seguridad de los mensajes, éstos deberían poder trocearse y ser enviados por diferentes caminos y medios, asegurándose la correcta recepción de los mismos al incluirse en cada paquete el destinatario, el remitente y el número de secuencia. Una vez recibidos estos paquetes por el destinatario del mensaje, el ordenador descartaba los duplicados y los reensamblaba de forma correcta. Si algún trozo del mensaje faltaba, automáticamente se solicitaba sin mayores problemas al estar éstos

[135] Rafael Chamorro, Ignacio Penedo. *Internet: estado del arte* [documento HTML]. Madrid: Asociación Profesional del Cuerpo Superior de Sistemas y Tecnologías de la Información de la Administración del Estado, [sin fecha]. Disponible en http://www.astic.es/estarte.htm (consultado el 4 de septiembre de 2000).

numerados. Esta idea revolucionaria dio lugar al desarrollo del protocolo de transmisión denominado NCP (*Network Control Protocol*)[136].

Así, en 1969 se creó la denominada red ARPANET, auténtica precursora de la posterior red Internet, siendo seleccionado como primer nodo de esta red el *Network Measurement Center* de la Universidad de California, en Los Ángeles, conectado en un principio al *Stanford Research Institute* (SRI), produciéndose el primer envío de mensaje de *host* a *host*. En poco tiempo se unieron a esta red dos nuevos nodos ubicados en la Universidad de California, en Santa Bárbara, y en la Universidad de Utah. Esta red fue creciendo paulatinamente durante los años 70 hasta llegar a conectar unos 100 ordenadores a principios de los 80. La siguiente figura ilustra el entramado o mapa lógico de esta red a principios de los años 70:

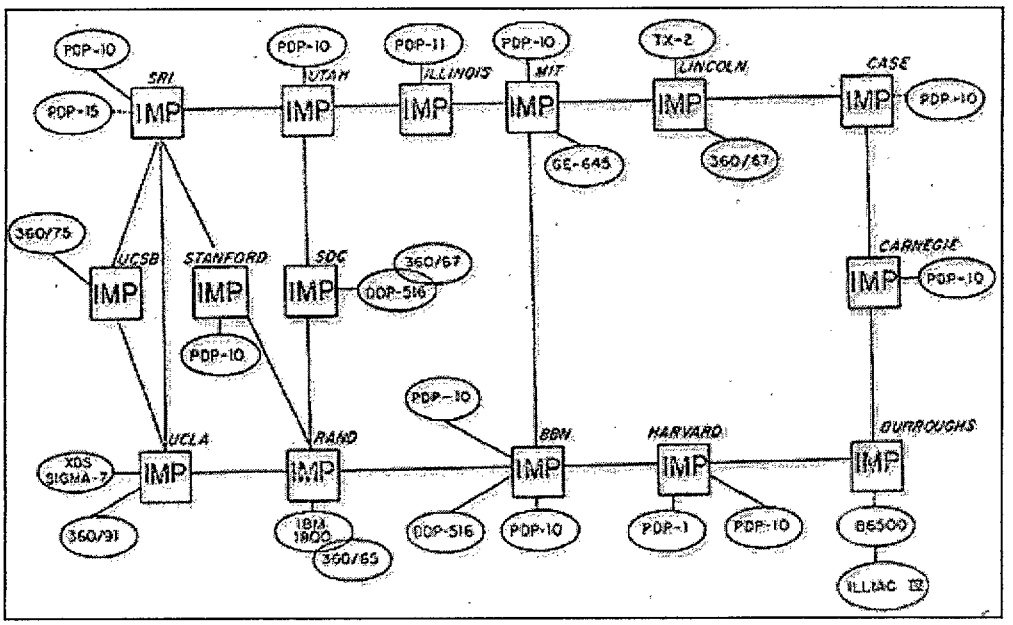


Figura I.9: Mapa lógico de la red ARPANET en abril de 1971. **Fuente:** *History of Internet and WWW*[137]

[136] Sergio Talens Oliag, José Hernández Orallo. *Internet: Redes de Computadoras y Sistemas de Información.* Madrid: Paraninfo, 1996, p. 4.

[137] Gregory R. Gromov. *History of Internet and WWW: The Roads and Crossroads of Internet History. Road 1. USA to Europe* [documento HTML]. Netvallley.com, 1995, rev. 2000. Disponible en http://www.netvalley.com/intval.html (consultado el 21 de agosto de 2000).

De igual forma, durante la década de los 70, muchos centros universitarios norteamericanos se empezaron a interesar por esta idea; esto es, interconectar sus centros de investigación a través de redes informáticas haciendo uso de innovadora filosofía de la conmutación de paquetes. Así, este testigo fue tomado por la *National Science Foundation* (NSF) estadounidense la cual pretendía dar acceso a sus importantes centros de supercomputación a otros investigadores de las universidades norteamericanas, por lo que nuevas universidades se fueron integrando dentro de ARPANET. Paralelamente a la evolución de esta red, fueron surgiendo nuevas redes informáticas, como BITNET y CSNET (*Computer Science Network*) que permitían interconectar otros centros universitarios y de investigación que habían quedado excluidos de la red ARPANET.

El surgimiento de diversas redes informáticas independientes dentro de los Estados Unidos y la necesidad de interconectar todas ellas para poder comunicar todos esos ordenadores entre sí hizo surgir el concepto de *internetting* y, como derivación del anterior, *Internetworking* (trabajo entre redes), conceptos de los cuales se derivaría posteriormente el término Internet. La idea era sencilla: construir una red de arquitectura abierta, en donde cada una de las subredes individuales podrían ser diseñadas y desarrolladas de forma aislada, con una interfaz propia para cada una de ellas y de acuerdo con un entorno tecnológico específico según los requerimientos de los usuarios de cada una de ellas; tan sólo se necesitaría un "lenguaje" común en el que pudieran comunicarse e intercambiarse datos todos los ordenadores de cada una de las redes[138]. Dos fueron los padres de tan importante idea: Robert Kahn y Vint Cerf. Así, debido a las limitaciones que presentaba el protocolo NCP de la red ARPANET para llevar a cabo este cometido, estos dos investigadores desarrollaron el protocolo conocido como TCP/IP (*Transmisión Control Protocol / Internet Protocol*). Según palabras del propio Cerf "necesitábamos encontrar un espacio de direcciones común que fuera independiente de las subredes. Diseñamos el protocolo de manera que el *host* estuviera aislado del interior de la red"[139]. Aunque las primeras especificaciones de este protocolo fueron presentadas en 1974 no sería hasta 1981

[138] B. M. Leiner... [*et al.*]. *Op. cit.*, p. 3.

[139] "Exploradores y pioneros (I)". *Global Communications*, nº 11, febrero de 1998, p. 69.

cuando quedase completamente definido, siendo adoptado formalmente por ARPANET como estándar en 1982 y sustituyendo, por tanto, al protocolo NCP.

Sin entrar en demasiados detalles técnicos sobre el protocolo TCP/IP, dado que ello escapa a las pretensiones de esta tesis, sí es necesario comentar que el término TCP/IP no hace referencia a una entidad única que combina dos protocolos, sino a un conjunto más grande y complejo de programas de software que proporciona servicios de red como terminal remoto, transferencias remotas de archivos y correo electrónico, entre otros[140]. La arquitectura en capas de los protocolos de Internet determina que unos se superpongan a otros. Sobre los propios de la red de transporte físico va situado en otro nivel el IP (*Internet Protocol*) y sobre él, casi siempre, los TCP (*Transmisión Control Protocol*). En un nivel superior están los protocolos de aplicación: Telnet que permite emulaciones de terminal remoto; FTP (*File Transfer Protocol*) que permite a un archivo de un sistema copiarse a otro sistema; SMTP (*Simple Mail Transfer Protocol*) utilizado para la transferencia del correo electrónico; NNTP (*Network News Transfer Protocol*) para la transferencia de servicios de noticias; Gopher para el acceso jerarquizado a la información; o, entre otros más, el HTTP (*HyperText Transfer Protocol*) para la comunicación de documentos hipertextuales en la WWWW. Todos estos protocolos de aplicaciones se apoyan en TCP/IP[141].

TCP es el protocolo de comunicaciones que proporciona una transferencia fiable de los datos y el protocolo IP es el responsable de mover los paquetes de datos (técnicamente conocidos como *datagramas IP*) ensamblados a través de la red. Estos dos últimos protocolos son los que otorgan el nombre al conjunto o familia de protocolos de Internet, lo que da idea de su importancia. Tal es así que, según palabras de D. E. Comer, el enorme éxito tecnológico de esta red radica precisamente en estos dos protocolos: "el protocolo IP permite que Internet incluya casi cualquier tipo de tecnología de comunicación de

[140] Tim Parker. *Aprendiendo TCP/IP en 14 días* [2ª ed.]. México [etc.]: Prentice-Hall Hispanoamericana, 1997, p. 33.

[141] J. T. Nogales Flores. *Op. cit.*, p. 145.

computadoras [...] El TCP resuelve los problemas de comunicación que el IP no puede solucionar y proporciona a las aplicaciones una comunicación confiable"[142].

Como señalábamos con anterioridad, todos los protocolos que se integran dentro de la familia TCP/IP se reparten dentro del mismo en cuatro capas o niveles[143]:

- Red: responsable de aceptar y transmitir los datagramas IP a través de una red concreta.
- Inter-red: nivel que gestiona la comunicación de una máquina a otra, aceptando paquetes del nivel de transporte y encapsulando éste en un datagrama IP junto a la información de control necesaria y determinando el camino que debe seguir para llegar a su destino. Este nivel también se encarga del control y la gestión de errores en la comunicación máquina a máquina.
- Transporte: es el nivel encargado de proporcionar la comunicación entre aplicaciones, regular el flujo de la información y asegurar que el transporte de la información se produzca de forma ordenada y sin errores.
- Aplicación: se trata del nivel de las aplicaciones de red, en el cual se utilizará al nivel transporte para el envío y recepción de los mensajes. Cada aplicación utilizará el estilo de transporte que mejor se adecue a sus necesidades, bien como secuencia de mensajes individuales o bien como un flujo continuo de información.

Destacaremos a continuación brevemente algunos de los hitos más importantes acaecidos en esta evolución de la red Internet:

- En 1983, la parte militar de ARPANET se separa de esta red, constituyendo la denominada MILNET, quedando Internet, por tanto, en manos de la agencia ARPA para otros usos en investigación avanzada fuera del campo militar. Es este año el que se considera como la fecha de nacimiento de la red Internet, con la constitución de la IAB (cuyo primer nombre fue el de *Internet Activities Board*) con la tarea de diseñar, construir y manejar Internet.

---

[142] Douglas E. Comer. *El Libro de Internet: todo lo que usted desea saber sobre redes de computadoras y acerca de cómo funciona Internet.* México D.F. [etc.]: Prentice Hall Hispanoamericana, 1995, p. 136.

[143] S. Talens, J. Hernández. *Op. cit.*, p. 55.

- En ese mismo año surge en Europa la EARN (*European Academic and Research Network*) que, aunque basada en los protocolos propietarios de la compañía IBM (RSCS/NJE), incluye la misma finalidad de interconexión de centros de investigación y académicos a través de redes informáticas para la transmisión de datos[144].
- A partir de mediados de los años 80 la red Internet comenzó a mundializarse, empezando a dispararse el número de usuarios (en especial el colectivo de estudiantes universitarios norteamericanos), proporcionando otros contenidos menos técnicos a esta red. Así, esta avalancha de nuevos usuarios conllevó la retirada de la agencia ARPA, pasando la red ARPANET a cargo de la NSF. Este organismo creará a su vez en 1986 la red NSFNET.
- Desde finales de la década de los 80 se van incorporando a Internet otras redes de países europeos (Francia, en 1988, fue el primer país), y de otras redes privadas de gran importancia, como BITNET (*Because It's Time Network*) y su correspondiente europea EARN. El gobierno norteamericano crea la NREN (*National Research and Education Network*) con la finalidad de fortalecer los servicios y recursos de información electrónica en la comunidad educativa.
- En enero de 1992 se anuncia la fundación de una organización dedicada a velar por el establecimiento de los estándares de Internet, la coordinación en la investigación en este campo y el desarrollo y crecimiento de esta red: la *Internet Society* (ISOC), organización de carácter mundial formada por usuarios, proveedores de acceso y fabricantes de hardware y software.
- En 1993 el presidente norteamericano William F. Clinton y el vicepresidente Albert Gore lanzan un ambicioso plan sobre política de información nacional difundido a través del documento *Technology for America's Economic Growth*, que sentará las bases para la confección y promulgación ese mismo año de la *U.S. National Information Infrastructure*

[144] Para una mayor información sobre el desarrollo de la red Internet en Europa, así como en nuestro país, recomendamos la consulta del artículo de Miguel A. Sanz. "Fundamentos históricos de la Internet en Europa y en España" [documento HTML]. *Boletín de RedIRIS*, nº 45, octubre 1998. Disponible en http://www.rediris.es/rediris/boletin/45/enfoque2.html (consultado el 4 de septiembre de 2000).

Act (NII)[145]. Este plan nacional en infraestructuras de información supuso, entre otras cosas, un espaldarazo definitivo al desarrollo de las ideas de Al Gore sobre la creación de las denominadas "autopistas de la información"[146].

- En 1994, y como consecuencia de la puesta en marcha del anterior plan, se eliminan en Estados Unidos todas las restricciones existentes al uso comercial de la red, dejando, asimismo, el gobierno estadounidense de controlar el flujo de información en Internet.
- En este año, la Comunidad Europea sienta las bases para el desarrollo de las nuevas tecnologías de la información en los países miembros a través del denominado "Informe Bangemann". El modelo europeo, con una visión más social que el modelo norteamericano, pretende hacer extensivo el uso de las tecnologías de la información y la comunicación a todos los sectores de la sociedad, desde la educación y el empleo hasta la implantación de servicios telemáticos en las administraciones públicas y en las PYMES[147].
- En octubre de ese mismo año se funda el *W3 Consortium* para el desarrollo de protocolos comunes que promuevan la correcta evolución de la Web[148].

---

[145] Para una mayor información sobre el impacto y repercusión que tuvo la *National Information Infrastructure* puede acudirser a Eva Mª Méndez Rodríguez. "Política del tándem Clinton-Gore en materia de información: El liderazgo de los Estados Unidos". En: Mercedes Caridad Sebastián (coord.). *La Sociedad de la Información: Política, Tecnología e Industria de los Contenidos.* Madrid: Centro de Estudios Ramón Areces: Universidad Carlos III de Madrid, 1999, pp. 4-36.

[146] La figura política de Albert Gore ha sido crucial en el desarrollo y promoción de la red Internet, como ponen de manifiesto Kahn y Cerf al señalarlo como "el primer líder político en reconocer la importancia de la Internet, así como en la promoción y soporte a su desarrollo". Para una más amplia información sobre este tema, véase Robert Kahn, Vinton Cerf. *Al Gore and the Internet* [documento HTML]. Internet Society. October 24, 2000. Disponible en http://www.isoc.org/internet/history/gore.shtml (consultado el 5 de febrero de 2001).

[147] Para una mayor información sobre las políticas europeas en materia de información y telecomunicaciones, puede consultarse el trabajo de Mercedes Caridad Sebastián. "Planes de la Unión Europea para alcanzar el próximo milenio en política del conocimiento". En: Mercedes Caridad Sebastián (coord.). *La Sociedad de la Información: Políticas, Tecnología e Industria de los Contenidos.* Madrid: Centro de Estudios Ramón Areces: Universidad Carlos III de Madrid, 1999, pp. 37-57.

[148] El origen y evolución de este consorcio será descrito con mayor profundidad en el siguiente apartado de este capítulo.

- Entre finales de 1994 y principios de 1995 el gobierno norteamericano pretende que la NII tenga un alcance mundial, transformándose para ello en la *Global Information Infrastructure* (GII). Este nuevo plan sentará las bases del desarrollo de la red Internet a una escala mundial e, indirectamente, sentando las bases para el liderazgo de los Estados Unidos sobre el sector de las telecomunicaciones.
- 1995 se considera como el verdadero momento de arranque de la red Internet como tecnología de uso orientada al gran público al proclamar el gobierno de los Estados Unidos, y por derivación de los objetivos marcados por la GII, que la red Internet puede perfectamente autocostearse, procediéndose a la disolución de la NSFNET. A partir de ese momento la gestión de las principales vías de acceso a Internet recaerá en manos de operadores de telecomunicaciones privados, los cuales ofrecerán precios y calidades variables de conectividad a la red.

La Internet es hoy en día un vasto conjunto de elementos tecnológicos interconectados y, sobre todo, de intereses de todo tipo, políticos, económicos, sociales y culturales. En todo este conjunto ciertamente complejo en donde tantos sectores convergen con intereses a menudo enfrentados, existen una serie de instituciones y organizaciones de carácter supranacional que tratan de dirigir y controlar el aparente caos reinante (tan sólo aparente) en esta red de redes[149].

A la cabeza de estas instituciones reguladoras se encuentra la ya mencionada **Internet Society** (ISOC)[150]. Esta sociedad, fundada en 1992, tiene como misión principal el fomento, desarrollo y mantenimiento de una serie de estándares para Internet, así como de las tecnologías que dan soporte a esta red, para "beneficio de todas las personas del mundo"[151]. Además de preocuparse de los aspectos más técnicos de la red funciona

[149] Para una mayor información al respecto de este punto, véase el artículo de David Martí. "¿Quién mueve los hilos en la red?". *Netmanía@*, nº 23, 1999, pp. 53-58.

[150] Toda la información oficial relativa a esta organización se encuentra disponible en la dirección http://www.isoc.org/

[151] *Internet Society Mission Statement* [documento HTML]. Internet Society, rev. 2001. Disponible en http://www.isoc.org/isoc/mission/index.shtml (consultado el 17 de enero de 2001).

también como elemento socializador en la universalización de Internet, promocionando programas educativos y de investigación en países con un menor desarrollo en infraestructuras y tecnologías de la información. Esta institución se estructura en órganos de gran peso e importancia, entre los que destacan los siguientes:

- La **Internet Architecture Board** (IAB)[152], fundada en 1983 y originalmente llamada *Internet Activities Board*, está constituida por una serie de veteranos de Internet con la misión de asesorar técnicamente a la ISOC en materia de desarrollo de estándares para la red Internet, tales como el TCP/IP, pero siempre desde postulados "generalistas con una buena visión global de todos los aspectos de la arquitectura de Internet"[153]. En gran medida, se puede decir que la IAB actúa como consejero técnico sobre temas relacionados con la arquitectura y con los procedimientos generales relacionados con Internet y su tecnología, desarrollando para ello un gran número de reuniones con sus brazos operativos: la IETF, la IRTF y la ISTF.
- La **Internet Engineering Task Force** (IETF)[154], creada en 1989 por la anterior institución, es el verdadero "brazo armado" de los aspectos más técnicos de la red Internet. Abierto en principio a la colaboración y participación voluntaria de cualquier persona interesada en la materia, tiene como misiones principales las cinco siguientes[155]:
  1. Identificar y proponer soluciones a los problemas técnicos y de funcionamiento más apremiantes en la Internet.
  2. Detallar el desarrollo o el uso de protocolos así como la arquitectura de primer nivel que permiten solucionar problemas técnicos de la Internet.
  3. Confeccionar recomendaciones para EL *Internet Engineering Steering Group* (IESG) relativos a la estandarización de protocolos y uso de los mismos en la Internet.

---

[152] Toda la información oficial relativa a este órgano se encuentra disponible en la dirección http://www.iab.org/.

[153] Brian Carpenter. *What Does the IAB Do, Anyway?* [documento HTML]. Internet Architecture Board, 1996. Disponible en http://www.iab.org/connexions.html (consultado el 17 de enero de 2001).

[154] Toda la información oficial relativa al IETF se encuentra disponible en la dirección http://www.ietf.org/

4. Facilitar la transferencia tecnológica desde la *Internet Research Task Force* (IRTF) al mayor número posible de personas que integran la comunidad de Internet.
5. Proporcionar un foro de intercambio de información dentro de la comunidad de Internet entre empresas comerciales, usuarios, investigadores, proveedores de acceso y gestores de red.

Esta importante institución es bien conocida dentro de la comunidad de usuarios de Internet por su labor en el desarrollo y aprobación de estándares a través de los denominados *Request for Comment* (RFC)[156].

- La **Internet Research Task Force** (IRTF)[157] creada en 1989, al mismo tiempo que la IETF, es la sección dedicada a la investigación tecnológica a largo plazo dentro del IAB, tratando temas diversos como recursos futuros, seguridad y privacidad en la red. Como se señala textualmente en el documento RFC 2014 que establece el funcionamiento de la IRTF, ésta "tiene la responsabilidad en la formación de grupos de investigación en materias relacionadas con los protocolos, aplicaciones y tecnologías de la Internet"[158].

---

[155] *The Tao of IETF – A Guide for New Attendees of the Internet Engineering Task Force* [documento HTML]. IETF Secretariat, [sin fecha]. Disponible en http://www.ietf.org/tao.html (consultado el 17 de enero de 2001).

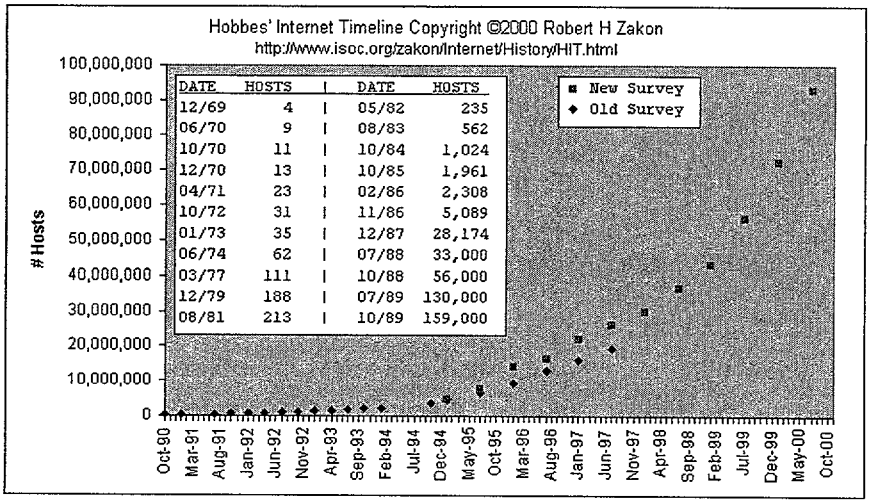
[156] El RFC 1310 de marzo de 1992, citado con anterioridad en este capítulo, estableció las pautas de procedimiento para la creación y documentación de estándares dentro de la IAB y de aplicación en la red Internet. Este RFC fue modificado por el RFC 1602 y, posteriormente, por el RFC 2026, vigente hasta nuestros días. Para una mayor información sobre este tema, véase el artículo de D. Crocker. *Making Standards: the IETF Way* [documento HTML]. Internet Society, 1993, rev. 2001. Disponible en http://www.isoc.org/internet/standards/papers/croker-on-standards.shtml (consultado el 17 de enero de 2001), así como el mencionado RFC 2026 en S. Bradner. *Request for Comments 2026: The Internet Standards Process* – Revision 3 [documento TXT]. Network Working Group, IAB, October 1996. Disponible en http://www.ietf.org/rfc/rfc2026.txt (consultado el 13 de noviembre de 2000).

[157] Toda la información oficial relativa a esta organización se encuentra disponible en la dirección http://www.irtf.org/

[158] A. Weinrib, J. Postel. *Request for Comments 2014: IRTF Research Group Guidelines and Procedures* [documento TXT]. Network Working Group, IAB, October 1996. Disponible en ftp://ftp.isi.edu/in-notes/rfc2014.txt (consultado el 17 de noviembre de 2000).

- La **Internet Societal Task Force** (ISTF)[159], es una de las instituciones relacionadas con la ISOC de más reciente creación (agosto de 1999), encaminada a tratar todos aquellos aspectos sociales y económicos del crecimiento de la red Internet que permitan llevar a buen puerto el precepto de la ISOC de una Internet para todos. Así, estudia y describe los factores que pueden determinar que en ciertas partes del mundo este desarrollo y crecimiento de la red no sea el adecuado.

El rápido avance y desarrollo de la red Internet por todo el mundo constituye, tal vez, uno de los hitos tecnológicos más importantes del siglo XX: transcurridas un par de décadas desde sus orígenes su implantación está muy extendida por el mundo. El siguiente gráfico ilustra perfectamente esta aceptación tecnológica, señalando la evolución sufrida a lo largo de esta última década en la puesta en marcha de ordenadores conectados a Internet:



DATE	HOSTS	DATE	HOSTS
12/69	4	05/82	235
06/70	9	08/83	562
10/70	11	10/84	1,024
12/70	13	10/85	1,961
04/71	23	02/86	2,308
10/72	31	11/86	5,089
01/73	35	12/87	28,174
06/74	62	07/88	33,000
03/77	111	10/88	56,000
12/79	188	07/89	130,000
08/81	213	10/89	159,000

Figura I.10: Evolución en el número de servidores en Internet. **Fuente:** R. H. Zakon. "Hobbes' Internet Timeline v5.2" [160]

[159] Toda la información oficial relativa a esta organización se encuentra disponible en la dirección http://www.istf.org/

[160] Robert H'obbes' Zakon. *Hobbes' Internet Timeline v5.2* [documento HTML]. Internet Society, rev. 19 November 2000. Disponible en http://www.isoc.org/guest/zakon/Internet/History/HIT.html (consultado

Reseñaremos brevemente que el imparable crecimiento en el número de usuarios que acceden a la red Internet, así como el vertiginoso crecimiento de contenidos informativos de todo tipo que circulan por dicha red informática, ha venido provocando en estos últimos años una fuerte saturación en las líneas de comunicación y en los equipos informáticos encargados de transmitir y controlar todo este inmenso flujo de información electrónica. Además, tal y como se comentó con anterioridad, la evolución sufrida por la Internet, desde una red orientada al intercambio de información entre investigadores y docentes universitarios de todo el mundo hacia un "espectáculo de masas" en donde el poder económico se ha convertido en uno de sus mejores aliados, hizo que a mediados de la década pasada algunas universidades y centros de investigación de los Estados Unidos se planteasen la construcción de una nueva Internet. Así, en octubre de 1996, representantes de 34 universidades norteamericanas junto con la *National Science Foundation* (NSF) lanzaron el proyecto denominado **Internet 2** (I2)[161], enmarcado dentro de la iniciativa de la administración del ex-presidente B. Clinton conocida bajo el nombre de *Next Generation Internet*. Este proyecto, al cual se unieron rápidamente otros socios empresariales y gubernamentales, pretende acelerar la próxima etapa del desarrollo de Internet con nuevos servicios y aplicaciones (en especial todas aquellas en las que se necesita una potente red de banda ancha, como la transmisión de vídeo y audio en tiempo real), aumentando para ello de forma espectacular las actuales capacidades que proporciona la Internet en la transmisión de datos[162].

La replica europea a este proyecto se denomina **TEN-155** (*Trans-European Networking at 155 Mbps*)[163]. Este nuevo prototipo de red IP, de características similares a la anterior, es

el 14 de febrero de 2001). En este documento electrónico, amén de detallar con exhaustividad los hitos más importantes en la evolución de la red Internet, se pueden encontrar un gran número de gráficos estadísticos sobre diversos aspectos técnicos de esta red.

[161] Toda la información oficial relativa a este desarrollo se encuentra disponible en la dirección http://www.internet2.edu/

[162] Laverna Saunders-McMaster. "Internet 2: An Overview of the Next Generation of the Internet". *Computers in Libraries*, v. 17, nº 3, March 1997, p. 57.

[163] Toda la información oficial relativa a este importante desarrollo se encuentra disponible en la dirección http://www.dante.net/ten-155/

resultado directo del proyecto europeo QUANTUM (*Quality Network Technology for User-Oriented Multi-media*).

Por último, no deseamos cerrar este punto sin hacer mención al fenómeno de la *Intranet*, tecnología informática que tanto ha dado que hablar en los últimos años. Muchas organizaciones de todo tipo han venido reconociendo que las tecnologías y los servicios de información derivados de la red Internet, y en especial de la *World Wide Web*, constituyen un modelo sólido, experimentado, fácil de introducir y de adaptar a nuevas necesidades. Este hecho llevó a muchas de estas entidades a considerar las posibilidades de adaptación de dichas tecnologías y servicios a un ámbito más reducido de trabajo, beneficiándose de este modo de todas las ventajas que así puede aportar a cada uno de sus usuarios, grupos o áreas de trabajo y, en definitiva, a la organización en su globalidad. La aplicación de dicha tecnología de red informática acotada a las necesidades y requerimientos internos de una organización es lo que se conoce con el término de Intranet. La **Intranet**, que no es más que una adaptación de elementos y sistemas de redes locales que ya existían con anterioridad a un nuevo modelo basado en la integración y homogeneidad de todas las herramientas y aplicaciones utilizadas a través del uso de unos protocolos y estándares de trabajo públicos y abiertos derivados de la familia del TCP/IP[164].

El servicio más importante que suministra hoy en día la red Internet, el servicio de la *World Wide Web*, que tan espectacular desarrollo ha tenido desde sus orígenes hasta nuestros días, será analizado con mayor detenimiento y profundidad en el primer capítulo de la siguiente Parte de esta tesis, dado el especial interés que tiene para la misma.

[164] Bonifacio Martín Galán. "La Intranet Documental y la Gestión de la Información en Entornos Corporativos". En: Carlos Olmeda Gómez (dir.). *Tesina. Doctorado en Documentación (programa 96-98)*. Getafe: Departamento de Biblioteconomía y Documentación, Universidad Carlos III de Madrid, septiembre de 1998, p. 3 (*mimeo*).

I.3.2. LA INFORMACIÓN JURÍDICA EN INTERNET

I.3.2.1. INTRODUCCIÓN Y CONTEXTO

La red Internet y sus servicios característicos han venido a potenciar de manera espectacular el concepto de la llamada *Sociedad de la Información*, la cual implica, entre otras muchas cosas, una sociedad en donde el uso de la información electrónica y las comunicaciones se convierten en la pieza central que mueve el desarrollo de las sociedades modernas. Dentro de toda esa maraña de información electrónica que circula por la red Internet, el acceso a la información jurídica por parte de los ciudadanos del mundo jugará un papel determinante en el desarrollo de dicha Sociedad de la información. Tal es su importancia que autores como A. D. Oliver-Lalana señalan que la red Internet ofrece en la actualidad al ciudadano un vasto espectro de posibilidades para obtener e intercambiar información jurídica, convirtiéndose en un tópico[165].

De manera general, se puede decir que existen tres clases de usos en la red Internet relacionados con la información jurídica; a saber[166]:

- *Comunicaciones electrónicas*: todas aquellas aplicaciones derivadas del uso de la red Internet que han ido siendo adoptadas por la práctica jurídica en el desarrollo de su labor, destacando especialmente el uso del correo electrónico y los servicios de videoconferencia.

[165] A. Daniel Oliver-Lalana. "Internet y el problema de la información jurídica del ciudadano: consideraciones en torno al acceso electrónico a la información jurídica". En: Miguel A. Davara Rodríguez (coord.). *Encuentro sobre Informática y Derecho* (14°. 2000-2001. Madrid). Madrid: ICADE, 2001, p. 340.

- *Publicación electrónica*: los mecanismos existentes en la red Internet para la publicación y difusión de información jurídica, en donde el servicio de la WWW ha sido (como se analizará con posterioridad) desde su aparición el gran escaparate donde mostrar todo tipo de contenidos.
- *Sistemas híbridos*: donde se combinan tanto los elementos propios de las comunicaciones como el de la publicación electrónica, y donde el resto de capacidades y servicios que la red Internet suministra hacen posible que la información jurídica alcance unas cotas de difusión difícilmente sospechadas hace varias décadas.

De igual modo, se pueden establecer de forma general tres componentes básicos en el método establecido para la difusión de la información jurídica en el espacio electrónico de la red Internet; éstos, según Y. Nobis, son[167]:

- *Transmisión de la información jurídica*: o el proceso de suministro de información jurídica desde las fuentes originarias de la misma. En la actualidad, existe un gran número de instituciones de todo tipo dedicadas a esta labor de difusión y suministro de información jurídica en Internet. Estas instituciones, públicas o privadas, suministran tanto información jurídica primaria (las normas y sentencias jurídicas en su estado original) como información secundaria, incluyendo aquí documentos de carácter doctrinal procedentes de distintos ámbitos del Derecho, revistas electrónicas de carácter jurídico, bases de datos referenciales o a texto completo, etc. Esta labor de transmisión jurídica ha sido vista tradicionalmente como una función esencial de las empresas editoriales, públicas o privadas, del ámbito del Derecho aunque, como se verá más adelante, esta situación ha cambiado notablemente con la irrupción de numerosos servicios jurídicos de todo tipo en el espacio electrónico de la WWW.

---

[166] Sanda Erdelez, Sheila O'Hare. "Legal Informatics: Applications of Information Technology in Law". En: Martha E. Williams (ed.). *Annual Review of Information Science and Technology*, v. 32. Medford, NJ: Information Today, 1997, p. 388.

[167] Yvonne Nobis. "Law in the Information Age". *Law Library*, v. 30, n° 1, 1999, p. 42.

- *Acceso y recuperación de la información jurídica*: en la actualidad existe una inmensa cantidad de información jurídica en Internet puesta a disposición del usuario pero esto no significa que su grado de accesibilidad sea el mismo siempre. En algunos casos, el acceso a esta información será simple y sencillo, además de gratuito, pero en otros muchos casos, especialmente en el acceso a bases de datos jurídicas comerciales disponibles en Internet, serán necesarios unos mayores conocimientos y formación para la recuperación de la información jurídica deseada. Además de esta disponibilidad y facilidad de acceso a esta información, de lo cual se hablará con posterioridad de forma más detallada, existen otros factores determinantes a la hora de recuperar información jurídica en Internet, como la fiabilidad de dicha información y las posibilidades de poner en relación unas informaciones con otras a través las capacidades hipertextuales que la Web nos proporciona. En el primer caso, resulta un factor determinante para cualquier jurista o interesado en el Derecho acceder y recuperar información jurídica con las suficientes garantías de veracidad del texto electrónico recuperado. El segundo caso, las cuestiones relativas a la ampliación de los contenidos informativos a través de las relaciones hipertextuales que se pudieran establecer dentro del documento, serán tratadas con mayor detenimiento en comentarios posteriores dentro de este punto.
- *Interpretación de la información jurídica*: el tratamiento y la interpretación de la información jurídica han venido siendo considerados, y justificados, como algo propio del terreno en el que se movían las empresas editoras de este tipo de información. El valor añadido que se le suministraba a la información jurídica primaria justificaba por sí mismo la existencia de este tipo de empresas. La adecuada organización a la que debe someterse toda información jurídica para facilitar su interpretación por juristas y estudiosos del Derecho ha impuesto que en el ámbito de la red Internet se reproduzcan y trasladen en muchos casos las técnicas y métodos de producción de bases de datos jurídicas por parte de estas empresas comerciales, añadiendo a todo ello las nuevas capacidades técnicas que la tecnología de Internet han sido capaces de suministrar. Pero, por otro lado, y debido a la facilidad que la red Internet presenta para la publicación de todo tipo de información electrónica, han venido apareciendo en los últimos años un gran número de servicios de información jurídica repartidos por todo el mundo. Tales tipo

de servicios ofrece en muchos casos una información jurídica gratuita pero escasamente tratada y organizada, sin unos mínimos mecanismos razonables para su búsqueda y localización, y con nulas garantías sobre la fiabilidad de la información jurídica publicada. Esta polémica entre suministradores de información jurídica comercial y gratuita, en Internet será tratada con posterioridad en este mismo punto del presente capítulo.

Analizaremos a continuación de forma más detallada algunas de las cuestiones principales aquí planteadas en relación con la difusión y el acceso a la información jurídica en Internet.

Sin ánimo de ser exhaustivos en nuestro análisis, resulta necesario, sin embargo, comentar ciertos aspectos de interés respecto al suministro de información de carácter jurídico en Internet, analizando en este primer momento aspectos tales como qué institución debe proporcionarla y el cobro o gratuidad en el acceso a la misma.

La expansión de la red Internet en sus primeras etapas, como ya se analizó en el primer punto del presente capítulo, estuvieron marcadas por un espíritu de amplia colaboración científica y, en gran medida, una ausencia de ánimo de lucro por parte de los servicios suministrados. Se trataba, en esencia, de un gran sistema electrónico mediante el cual diferentes instituciones de todo el mundo procedentes principalmente del ámbito de la investigación científica y la docencia universitaria podían intercambiar experiencias profesionales e información de todo tipo. En muchos casos, los bancos de datos de estas instituciones se ponían a libre disposición del resto de usuarios con el fin de facilitar y propagar estas ideas de cooperación científica. Esta misma filosofía fue rápidamente extendiéndose a otros sectores de la sociedad, impulsados por la apuesta firme que la mayoría de los gobiernos de países desarrollados hacían por hacer realidad una Sociedad de la información en donde el suministro de información electrónica generada por sus distintas Administraciones públicas tuviese las suficientes garantías de libertad de acceso y gratuidad para el ciudadano.

En el caso de la información jurídica de carácter público en formato electrónico las opiniones y corrientes han venido siendo diversas y controvertidas pues diferentes han sido los intereses que se han movido en torno a la misma. La raíz del problema reside en la

capacidad y potestad que tienen los diversos interesados para distribuir información jurídica en este espacio electrónico.

Como acertadamente señala T. R. Bruce, del Instituto Jurídico de la Escuela de Derecho de la Universidad de Cornell, cada proveedor de información jurídica de carácter público se sitúa, por definición, entre el creador o conjunto de creadores de esta información y el consumidor o grupo de consumidores de la misma. De este modo, cada uno de los distintos proveedores asumirá unas determinadas características dependiendo del valor de la información jurídica que vaya a suministrar y de sus potenciales consumidores[168]. Siendo esto cierto, cabe considerar quién ha de ejercer dicho papel de suministrador de la información jurídica que emana de los poderes públicos del Estado. Este mismo jurista plantea esta disyuntiva en las distintas alternativas existentes; a saber[169]:

- El sector público frente al sector privado, de forma general.
- Publicación por parte del Gobierno frente a la publicación por otras partes, tanto del sector privado como del sector público.
- Publicación por parte de un organismo central frente a la publicación directa por parte de las instituciones que crean esta información.

En ese primer estadio de debate entre el sector público y el sector privado como transmisión de información jurídica de carácter público dentro de la red Internet entran en juego diversos factores, en especial de índole política. Los diversos modelos de economía política que están establecidos a lo largo y ancho del planeta establecen una mayor o menor intervención del sector público en los todos los asuntos relativos al funcionamiento de los países. Desde la óptica de un modelo intervencionista por parte del sector público el control y difusión de esta información debe, por tanto, ser realizado desde este sector,

---

[168] Thomas R. Bruce. "Public Legal Information: Focus and Future" [documento HTML]. *The Journal of Information, Law and Technology* (JILT), nº 1, 2000. Disponible en http://www.law.warwick.ac.uk/jilt/00-1/bruce.html (consultado el 15 de noviembre de 2000).

[169] Thomas R. Bruce. *Some Thoughts on the Constitution of Public Legal Information Providers* [documento HTML]. Ithaca, NY: Cornell Law School, Cornell University, 2000. Disponible en http://www4.law.cornell.edu/working-papers/open/bruce/warwick.html (consultado el 10 de enero de 2001).

siendo además este acceso libre y gratuito para los ciudadanos del país. Por el contrario, en los sistemas más liberales se deja que el sector privado sea quien de forma especial se ocupe de la difusión y comercialización de la información jurídica. Muchos han sido los argumentos a favor y en contra de una u otra postura.

En un seminario celebrado en 1998 en la Universidad británica de Warwick con la idea de debatir estos asuntos[170], se reunió a diferentes personalidades representantes de cada uno de los sectores en conflicto (del sector privado, del ámbito público, del sector universitario y del sector de las bibliotecas) para que expusieran sus planteamientos. Desde el sector privado se justificaba su labor incidiendo sobremanera en la necesidad e importancia del establecimiento de fronteras claras de actuación entre los servicios que proporcionan un acceso gratuito a la información jurídica y aquellos proveedores comerciales de bases de datos jurídicas en Internet, aunque reconociendo que la gran dificultad de esta tarea estriba en decidir qué productos informan al ciudadano y cuáles ayudan al jurista; esta frontera debería estar marcada, por tanto, en el valor añadido que se le proporciona a las fuentes primarias del Derecho. Sin embargo, desde el sector público se incide en el hecho de que, de forma general, la información jurídica debería suministrarse de forma gratuita basándose en una serie de principios democráticos como la obligación del conocimiento de la ley por parte de los ciudadanos o los propios derechos de éstos a estar informados, especialmente considerando la complejidad y el constante aumento de normas que les afectan; se señala también que la información jurídica primaria, creada en foros públicos, no es propiedad de ningún gobierno, y finalmente, la obligación que tienen los estados democráticos de suministrar a los ciudadanos información pública que les puede afectar.

Aún más, desde muchos sectores de la sociedad, incluido el poder judicial, se ha venido señalando la paradoja existente en la mayoría de países del mundo en cuanto a que si la información jurídica es generada desde diversos organismos y dependencias del Estado, y suministrada a las casas editoriales privadas, éstas posteriormente vendan la misma información a estos mismos organismos de la Administración. Esta extraña paradoja ha

[170] Kelsie Aquatias. "New Directions in Legal Information Systems. Conference Report" [documento HTML]. *The Journal of Information, Law and Technology* (JILT), nº 2, 1998. Disponible en http://elj.warwick.ac.uk/jilt/confs/98_2cti/cit.htm (consultado el 24 de enero de 2000).

hecho que algunos países en los que se están atravesando graves crisis económicas cancelen sus suscripciones a las bases de datos jurídicas comerciales y opten por desarrollar sus propias bases de datos de forma cooperativa o, en otros casos, por recurrir a la información jurídica que existe actualmente en Internet de forma gratuita[171].

Sin embargo, y a pesar de estos razonamientos democráticos, existen una serie de impedimentos que tradicionalmente han obstaculizado este acceso público y gratuito a la información jurídica en muchos países de nuestro entorno socio-político, señalados por T. Bruce en los siguientes[172]: falta de financiación, contratos de explotación en exclusividad por parte de las editoriales jurídicas o derechos de reproducción que impiden su difusión pública y gratuita. De igual modo, se ha venido esgrimiendo desde el sector privado que esta labor de difusión de información jurídica se realiza de forma más conveniente desde este sector dado que existen los mecanismos para que dicha información se encuentre perfectamente actualizada, sea más completa, esté analizada y organizada de forma rigurosa y se distribuya de mejor manera entre los profesionales que han de hacer uso de ella[173].

Al margen de esta polémica, surgen de igual modo diversas opiniones sobre cuál es la institución idónea (dejando al margen a las editoriales comerciales, bien sean públicas o privadas) para la difusión de información jurídica de interés. Los principales debates a este respecto han surgido bien del ámbito académico o de los centros de documentación y bibliotecas especializadas en documentación jurídica. Desde estos ámbitos de actuación, y en especial desde los Estados Unidos, se descarta que, en principio, sean los propios organismos creadores de la información jurídica (parlamentos, tribunales, etc.) quienes proporcionen servicios de valor añadido a la información que producen dado que esta no

[171] Tal es el caso, por ejemplo, de Argentina donde diversos Magistrados y Jueces han planteado esta compleja situación en su relación con los proveedores comerciales de bases de datos jurídicas. Para una más amplia información, véase en la dirección http://www.diariojudicial.com/reportajes.asp?ID=7906 (consultado el 2 de noviembre de 2001).

[172] T. R. Bruce. "Public Legal...". *Op. cit.*, http://www.law.warwick.ac.uk/jilt/00-1/bruce.html

[173] Richard A. Danner. *Dissemination of Legal Information: Social and Political Issues in the United States.* Durham, NC: Duke University School of Law, 1998. Disponible en http://www.law.duke.edu/fac/danner/adij.htm (consultado el 8 de febrero de 2001).

es la función que tienen encomendada. Se señala, pues, que un lugar natural para llevar a cabo esta labor está en las Escuelas y Facultades de Derecho del entorno universitario pues es aquí donde se están efectuando las principales investigaciones para el tratamiento de la información y la documentación jurídica, así como la definición y promoción de estándares abiertos para la descripción de la información contenida en los documentos electrónicos[174]. De igual modo, los especialistas en información y documentación jurídica, atendiendo al hecho de que la presencia de sitios web de bibliotecas jurídicas se ha incrementado espectacularmente en estos últimos años[175], reclaman el reconocimiento de su importancia en dicha tarea pues tan sólo se trata de extender las técnicas y procesos que tradicionalmente se han venido llevando a cabo en estas instituciones en el tratamiento de la información a este nuevo entorno electrónico de desarrollo de colecciones[176].

Sea como fuere, lo que sí resulta un hecho cierto en nuestros días es la especial sensibilización que los Gobiernos de todo el mundo están demostrando en cuanto al uso de las tecnologías de la red Internet para suministrar información de carácter público a los ciudadanos. Son innumerables los países del mundo en donde sus gobiernos centrales o autonómicos han desarrollo o han empezado a desarrollar en fechas recientes auténticos *portales web* con información sobre las distintas Administraciones y Servicios públicos, ofreciendo al ciudadano un espacio centralizado para la localización de la información deseada o, en muchos casos, para establecer un nuevo mecanismo de comunicación y relación (electrónica) con dichas Administraciones[177]. Este tipo de desarrollos públicos se

[174] T. R. Bruce. "Public Legal...". *Op. cit.*, http://www.law.warwick.ac.uk/jilt/00-1/bruce.html

[175] Resultan ciertamente reveladores los datos ofrecidos en el artículo elaborado por R. C. Vreeland al respecto de este tema. Para una más amplia información, véase Robert C. Vreeland. "Law Libraries in Hyperspace: A Citation Analysis of Wolrd Wide Web Sites". *Law Library Journal*, v. 92, n° 1, 2000, pp. 9-25.

[176] John P. Joergensen. "Are Non-Profit Internet Publishers the Future of Legal Information?". *Legal Reference Service Quarterly*, v. 17, n° 1-2, 1999, p. 35.

[177] Sirva como ejemplo, lo sucedido en nuestro país con el *portal* que la Administración General del Estado ha puesto en marcha en fechas recientes. Aunque aún no está desarrollado de forma completa pues los servicios ofrecidos a través del mismo distan mucho de lo que se está realizando en otros países de nuestro entorno, desde aquí es posible acceder de forma centralizada a gran parte de la información que se suministra en

basa en los potenciales beneficios que para los distintos Gobiernos del mundo ofrece la WWW, perfectamente sintetizados por el *Center for Technology in Government* de la Universidad de Albany en los siguientes[178]: ayuda a los Gobiernos a expandir y enriquecer su audiencia; ofrece a los ciudadanos, a cualquier hora y en cualquier lugar acceso a la información gubernamental; proporciona un punto central de entrada de datos procedentes de diversas fuentes; da una visión de consumo de información a los ciudadanos; e incrementa el impacto de esta información con el uso de imágenes, audio y vídeo.

Finalmente, señalaremos brevemente, pues será motivo de desarrollo a lo largo de esta tesis doctoral, algunas de las características técnicas que el espacio de la *World Wide Web* proporciona a los documentos electrónicos que en él residen.

S. Erdelez y S. O'Hare señalan que el uso de la Web constituye desde su propia filosofía de construcción una manifestación perfecta de lo que es en realidad el entramado de información y datos relacionados que caracteriza a los documentos jurídicos[179]. En los medios impresos tradicionales e, incluso, en los primeros sistemas de acceso a los textos electrónicos, el jurista tenía que ir de un modo lineal de un texto a otro para obtener una información completa de lo que deseaba localizar. Sin embargo, con la concepción en el modo de funcionamiento de la Web y la materialización en la construcción de documentos hipertextuales basado en su lenguaje más característico, el HTML, del cual se hablará detenidamente en el apartado III.1.2 del presente trabajo, los documentos no están ya limitados por las dos dimensiones de la hoja de papel. Este nuevo entorno de consulta y lectura no secuencial de la información permite al usuario moverse o *saltar* desde un concepto existente en un documento electrónico a otro concepto relacionado, dentro del mismo documento o a otro distinto, formando de este modo una maraña de enlaces

diversos organismos y dependencias de la Administración Central, así como a la consulta de un gran número de bases de datos de todo tipo. Para una más amplia información, véase http://www.administracion.es/

[178] Información disponible en http://www.ctg.albany.edu/projects/inettb/univ/Reports/CH2/overview.html

[179] S. Erdelez, S. O'Hare. *Op. cit.*, p. 389.

interconectados que ponen en estrecha relación informaciones y documentos vinculados lógicamente.

Con respecto a las ventajas que la tecnología Web y los lenguajes de marcado de los documentos que residen en este espacio de publicación (HTML, y más importante aún, XML) aportan a los documentos jurídicos y, por extensión, a la construcción y diseminación de esta información contenida en bases de datos legislativas y jurisprudenciales, J. T. Nogales y M. C. Arellano señalan las siguientes[180]:

- *Tecnología de gran implantación*: Los navegadores o *browsers* de web están implantados en cualquier ordenador con acceso a Internet y HTML se ha convertido en un formato de intercambio universal de documentos textuales. Igualmente, existen innumerables programas, muchos de ellos gratuitos, tanto para el establecimiento de servicios de información en este espacio electrónico como para el acceso a los mismos.
- *Documentos de cualquier tamaño y con integración de medios diversos*: Los documentos electrónicos en este entorno se caracterizan por la inclusión de medios diversos, como texto, tablas, gráficos o dibujos, sonidos, vídeo, etc., pudiendo ser éstos de cualquier tamaño.
- *Capacidad hipertextual*: Ello permite integrar los enlaces de las referencias en el propio texto, tanto las internas al documento como las externas a otros documentos, capacidad ésta que en el caso de los documentos jurídicos es extremadamente importante por las relaciones explícitas que se dan en el texto de dichos documentos.
- *Diversidad de soportes y medios de difusión para una misma base de datos*: Sin tener que establecer modificación alguna, el conjunto de documentos hipertextuales puede ser difundido a través de Internet por medio de un servidor web o bien, en la forma tangible del CD-ROM u otros soportes ópticos y magnéticos. Sobre este conjunto documental que conforma la base de datos es posible integrar fácilmente un motor de búsqueda que permita realizar las funciones clásicas de búsqueda de información precisa, en especial si se utilizan las últimas tecnologías aplicadas a la Web, caso del metalenguaje XML.

A lo largo de esta tesis doctoral defenderemos la utilidad real que tienen estas tecnologías de la Web para el tratamiento de la información jurídica y construcción de bases de datos que han de ser servidas en Internet, y de forma más concreta el último de los lenguajes de marcado de documentos electrónicos establecido por el Consorcio de la Web (W3C): el metalenguaje XML.

Pero antes de dar paso al significado y alcance de los lenguajes de marcado de texto, así como al surgimiento y evolución sufridos por éstos hasta llegar al mencionado XML, resulta conveniente ofrecer, aunque sea de forma breve, algunos de los ejemplos más significativos de información jurídica disponible a través de Internet en la actualidad.

I.3.2.2. PRINCIPALES SERVICIOS DE INFORMACIÓN JURÍDICA EN INTERNET

Las tecnologías de la información en estos últimos años, han revolucionado notablemente el modo en el que la información jurídica es diseminada por los órganos judiciales, legislativos y las agencias gubernamentales, así como los métodos de acceso a dicha información por todos los ciudadanos.

El número de sitios web existentes en Internet que suministran información legislativa o jurisprudencial ha crecido espectacularmente en estos últimos años. Por lo general, serán los distintos órganos de la Administración Pública, en especial de la Administración de Justicia, los que pongan a disposición de los ciudadanos las fuentes primarias del Derecho, esto es, las normas legislativas y la jurisprudencia de los distintos Tribunales. Las fuentes

[180] J. T. Nogales Flores, M. C. Arellano Pardo. *Op. cit.*, p. 181.

secundarias se encuentran igualmente disponibles en Internet, incluyendo directorios de profesionales del campo del Derecho, enciclopedias y diccionarios, formularios, periódicos y revistas, y, por supuesto, bases de datos jurídicas.

Resulta necesario comentar que esta proliferación de sitios web con información jurídica emanada de los poderes del Estado y puesta ha disposición de manera libre y gratuita para el ciudadano, ha tenido su base en las manifestaciones e iniciativas de reivindicación que diversos colectivos de todo el mundo han venido protagonizando en estos últimos años. Baste destacar lo sucedido en España hace ya unos cuantos años, en donde la iniciativa conocida por el nombre de *¡BOE Gratis, ya!*, con la finalidad de que dicha institución editorial de la Administración pública pusiera a disposición de todos los ciudadanos un acceso gratuito a través de Internet a los contenidos del Boletín Oficial del Estado. En esta iniciativa colaboraron instituciones tan importantes como el Il.lustre Col.legi d'Advocats de Girona, el Colegio de Abogados de Alicante, el Colegio de Graduados Sociales de Valencia, la sección andaluza y aragonesa de la Internet Society, la Confederación Nacional de Sordos, ANPE (Sindicato Independiente), ISTAS (de CC.OO.), la Liga Española de Asociaciones de CB y Radio, la Unión Romaní, APFJA (Asociación de funcionarios), AEMAR (Asociación empresariales) y la Unión de Consumidores y Usuarios. Con la recogida de más de 10.000 firmas de adhesión a esta iniciativa y la presión ejercida desde diversos medios de comunicación al dar amplio seguimiento a esta propuesta popular, la Comisión del Régimen Jurídico de las Administraciones Publicas del Congreso de los Diputados aprobaría el 11 de mayo de 1999 la consulta gratuita del BOE a través de Internet[181]. El acceso a los contenidos del BOE sería gratuito pero los servicios de información jurídica contenida en sus bases de datos seguirían teniendo un carácter comercial dado que se les considera servicios con valor añadido.

Desde entonces y hasta nuestros días han sido numerosos los Boletines oficiales de otras instituciones y otros gobiernos autonómicos y provinciales de nuestro país que han ido poniendo igualmente la información jurídica contenida a disposición de todos los

[181] Información aparecida en mayo de ese mismo año en el antiguo portal jurídico en Internet Derecho.org, en la dirección http://derecho.org/boe / A partir de ese momento el BOE tendría que dar acceso de forma gratuita al Boletín Oficial a través de su sitio web (http://www.boe.es/)

ciudadanos de forma gratuita a través de Internet. Tal es el caso, por ejemplo del Senado (http://www.senado.es/), el Congreso de los Diputados (http://www.congreso.es/), el Boletín Oficial de la Junta de Andalucía (http://www.junta-andalucia.es/fr_boja.htm), el Boletín Oficial de Aragón (http://www.aragob.es/sid/boaboa.htm), el Boletín Oficial de Canarias (http://www.gobcan.es/boc/), el Diario Oficial de Castilla y León (http://www.jcyl.es/bocyl/), el Diari Oficial de la Generalitat de Catalunya (http://www.gencat.es/diari/index.htm), el Boletín Oficial del País Vasco (http://www.euskadi.net/castellano/bopv/welcome.html), el Diario Oficial de Galicia (http://www.xunta.es/doga/index.htm), el Boletín Oficial de la Comunidad de Madrid (http://www.comadrid.es/bocm/) y otros tantos de cada una de las diversas Comunidades Autónomas que se integran dentro del Estado español.

Centrándonos en el caso de las bases de datos jurídicas existentes en Internet, nacionales o extranjeras, es necesario establecer las oportunas diferencias entre lo que es un *sitio* jurídico y una *base de datos* jurídica dentro de este espacio electrónico. Siguiendo a M. Roznovschi, un sitio jurídico en Internet tiene una dirección electrónica, o URL, claramente definida y están integrados por un cierto número de documentos HTML, una colección de datos o simplemente un índice de fuentes de interés recopiladas por un especialista en la materia. Por el contrario, las bases de datos jurídicas están compuestas por una serie de datos o textos jurídicos a los cuales se les ha sometido a un proceso de descripción y análisis documental, almacenados en un servidor por alguna institución u organismo, público o privado, con un software capaz de establecer diversos mecanismos para la búsqueda y recuperación de la información y, finalmente, con unos criterios estables de actualización periódica de la información jurídica contenida[182]. Desde esta perspectiva, muchos de los actuales sitios en Internet con contenidos jurídicos no se puede decir con exactitud que incorporen bases de datos entre los servicios prestados, como veremos con posterioridad, pues se trata únicamente de recopilación de enlaces de interés para los

[182] Mirela Roznovschi. *Evaluating Foreign and International Legal Databases on the Internet* [documento HTML]. Law Library Resource Xchange (LLRX.com), March 1, 1999. Disponible en http://www.llrx.com/features/evaluating.htm (consultado el 21 de febrero de 2001).

profesionales del Derecho o un acceso a la legislación y/o la jurisprudencia nacional o internacional a través de simples listados.

En cualquier caso, estas barreras tienden a romperse pues, como demostraremos en esta tesis, una colección de documentos marcados con HTML o, mejor aún, con XML, junto con la incorporación de un motor de búsqueda ad hoc de suficiente potencia, constituye una alternativa plausible a las bases de datos documentales convencionales.

Debido a la gran cantidad de sitios web de interés jurídico, así como a las diversas bases de datos jurídicas puestas a disposición de los usuarios de la Web, tanto por instituciones y organismos públicos como por empresas editoras, resulta del todo imposible hacer una exposición exhaustiva de todos estos recursos (aún simplemente anunciarlos) sin que con ello se vea prolongado en exceso el contenido de esta tesis doctoral[183]. Por este hecho, tan sólo anunciaremos aquí los grandes sistemas de búsqueda de información en la Web, tanto los de contenido general como aquellos que están especializados en información jurídica y las distintas bases de datos jurisprudenciales de organismos específicos que se encuentran disponibles en Internet. Desde los grandes servicios de información jurídica en Internet es posible acceder a otros recursos electrónicos específicos en cada uno de los campos del Derecho (legislación, jurisprudencia y doctrina), tanto de ámbito nacional como internacional. De igual modo, desde este tipo de recursos generales es posible acceder a las direcciones electrónicas de aquellas instituciones y organismos oficiales que han venido manteniendo y poniendo a disposición del ciudadano información de interés jurídico y, en algunos casos, bases de datos con información. Los recursos que se citan a continuación

[183] Además de los recursos electrónicos relatados a continuación, resultan de gran interés tres obras impresas sobre esta cuestión, a las cuales remitimos de igual modo para una más amplia información sobre este tipo de recursos en Internet; éstas son:

- Miguel Ángel Gonzalo Rozas. "La Documentación Jurídica en Internet". En: Mateo Maciá (ed.). *Manual de Documentación Jurídica.* Madrid: Síntesis, 1998, pp. 447-495.
- Serge Guinchard, Michèle Harichaux, Renaud de Tourdonnet. *Internet pour le Droit: Conexion – Recherche – Droit.* Paris: Montchrestien, 1999.
- Ivonne J. Chandler "Legal Information on the Internet". *Journal of Library Administration*, v. 30, n° 1-2, 2000, pp. 157-208.

constituyen, por tanto, un punto inicial desde el cual partir para localizar la información de carácter jurídico que se desea.

• MOTORES DE BÚSQUEDA, DIRECTORIOS Y METABUSCADORES GENERALES

En cuanto a los recursos existentes en la Web con información jurídica, es necesario destacar en primer lugar los grandes buscadores de información general en Internet, grandes repositorios de documentos electrónicos, en donde, lógicamente, podrán encontrarse recopiladas todas aquellas páginas web con información jurídica. Dentro de esta gran categoría es posible distinguir entre **motores de búsqueda** (recursos recopilados de forma automática por un software y almacenados sin ningún criterio sistemático, permitiendo la búsqueda a través de una o varias palabras relevantes, fechas, etc.) y **directorios** (recursos recopilados por personal especialista en diversos campos y en donde el almacenamiento y acceso a los mismos se realiza principalmente a través de la navegación o *browsing* de una estructura jerarquizada según diversas materias del conocimiento humano). La frontera entre ellos es, en muchas ocasiones, borrosa pues ambos modelos suelen asociarse para incluir la potencialidad del motor de búsqueda por palabras junto a una estructuración temática de sus contenidos.

En el primer caso nos encontramos sistemas tan importantes y utilizados por los usuarios de la red como son *Google*, *AltaVista*, *HotBot*, *Excite*[184], etc. Entre los segundos, caben destacar como ejemplos representativos a *Yahoo!*, *Galaxy*, *Argus Clearinghouse* o, realizados desde el ámbito universitario y científico, *BUBL* o *The World Wide Web Virtual Library* [185].

[184] Sus direcciones electrónicas son, respectivamente, http://www.google.com/, http://www.altavista.com/, http://www.hotbot.com/, http://www.excite.com/

[185] Sus direcciones electrónicas son, respectivamente, http://www.yahoo.com/, http://galaxy.einet.com/, http://www.clearinghouse.net/, http://bubl.ac.uk/ y http://www.vlib.org/

Destacar igualmente, la existencia de otros servicios en la Web denominados **metabuscadores**, encargados de lanzar una misma sentencia de búsqueda a diversos buscadores específicos. Ejemplos significativos de este tipo de servicios lo constituyen *MetaCrawler* o *One2Seek*[186].

- **PORTALES Y BUSCADORES INTERNACIONALES DE INFORMACIÓN JURÍDICA**

Se trata en este caso de buscadores (motores de búsqueda y directorios) especializados en el campo del Derecho, almacenando y poniendo a disposición de los usuarios enlaces a recursos en la Web con información exclusivamente jurídica. En muchos casos, al igual que les sucede al anterior grupo, han evolucionado hacia auténticos portales de servicios web (además de la recopilación de enlaces de interés jurídico ofrecen otra serie de valores añadidos, como cuentas de correo electrónico, información política, cultural, social, etc.). Dentro de estos grandes buscadores internacionales de información jurídica resultan destacables los siguientes:

- **FindLaw** (http://www.findlaw.com/): Uno de los sitios fundamentales para la localización de información jurídica existente en Internet. Estructurado de un modo clásico, permite localizar la información de interés tanto a través del motor de búsqueda por palabras clave navegando a través de un directorio temático con clasificación jerárquica, dividiendo el conjunto de recursos almacenados en cuatro grandes bloques: los que pueden ser de interés para los profesionales del Derecho, los de interés para los estudiantes universitarios en este campo, para los profesionales de los negocios y del comercio y, finalmente, para el público en general. Tal es la importancia de este portal jurídico que, como se verá en capítulos posteriores, ha venido patrocinando y financiando económicamente algunos proyectos de investigación y a las organizaciones

[186] http://www.metacrawler.com/ y http://one2seek.com/, respectivamente.

que los llevan a cabo en el campo del tratamiento informático de la información jurídica. Esta compañía adquirió hace unos años algunos de los servicios web más importantes en materia jurídica, como han sido el *Cyberspace Law Center* (actualmente en la dirección http://cyber.findlaw.com/), de especial utilidad para los estudiosos de los problemas jurídicos del ciberespacio y los nuevos delitos en este terreno, el directorio temático, y *LawCrawler* (http://lawcrawler.findlaw.com/), directorio tradicional de recursos jurídicos en la web procedentes de todo el mundo.

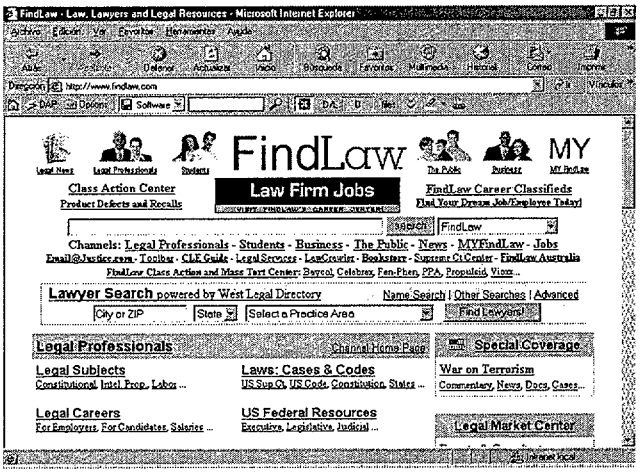


Figura I.11: FindLaw. Portal jurídico en Internet.

- **Guide to Law Online** (http://www.loc.gov/law/guide/index.html): Importante y prestigiosa guía de recursos de información jurídica en Internet elaborada por la *Law Library of Congress* de los Estados Unidos como parte del ambicioso proyecto GLIN (*Global Legal Information Network*) (http://www.loc.gov/law/glin/index.html), para la creación de una magna base de datos de legislación, jurisprudencia y otros documentos jurídicos de interés de todo el mundo. Se trata de una guía de recursos web de interés jurídico en donde principalmente se recopilan enlaces de sitios web de los Estados Unidos aunque cuenta con un apartado de ámbito internacional, así como un acceso a través de materias jurídicas.

- **Internet Legal Resource Guide** (http://www.ilrg.com/): Gran índice de recursos jurídicos en Internet en donde se han seleccionado más de 4.000 sitios web de unos 240 países del mundo. Además, almacena de forma local más de 800 páginas web y otro tipo de ficheros electrónicos considerados de especial interés por su contenido jurídico. Dado que está desarrollado en los Estados Unidos, la mayoría de los recursos citados se encuentran ubicados dentro de dicho país. Para el acceso a los recursos jurídicos descritos puede emplearse bien el motor de búsqueda por palabras clave o la estructuración temática de los contenidos (en secciones tales como información académica, información para los profesionales del Derecho, investigación jurídica en los Estados Unidos, investigación en el resto de países del mundo, así como otras cuestiones relativas a fuentes de información y noticias en Internet).

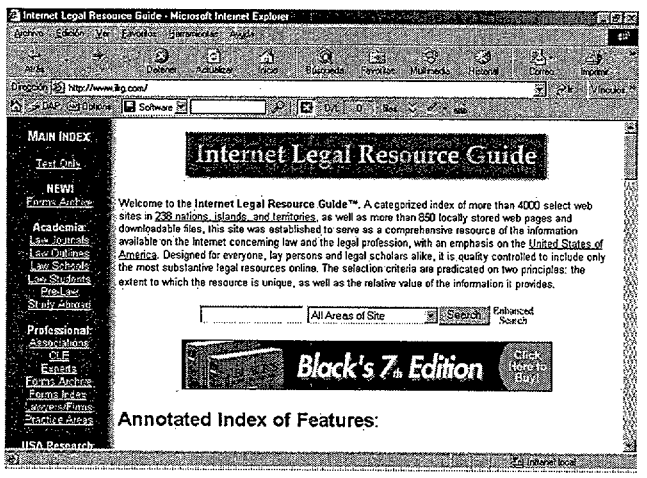


Figura I.12: Internet Legal Resource Guide. Servicio de Información jurídica en Internet.

- **HierosGamos – Legal Research Center** (http://www.hg.org/): Desarrollado en los Estados Unidos, por lo que su principal cobertura de recursos web de interés se situará en este país, abarca igualmente un gran número de enlaces a información jurídica de todo el mundo sobre múltiples aspectos (legislación, asociaciones profesionales, centros de noticias, centros de estudio, negocios, etc.). También recoge una clasificación geográfica de sus recursos electrónicos, entre ellos los españoles.

- **CataLaw** (http://www.catalaw.com/): Se trata de un catálogo de catálogos de sitios web de información jurídica (metadirectorio) desde el cual podemos acceder a los que sean de nuestro interés según su orientación temática, su situación o cobertura geográfica o por otros criterios preestablecidos por este servicio (bibliotecas jurídicas, editores especializados, periódicos y revistas jurídicas, escuelas de Derecho, etc.).

- **Law Guru – Legal Research Tool** (http://www.lawguru.com/multisearch/multimenu.html): Se trata de un claro ejemplo de metabuscador aplicado al campo de la información jurídica al permitir lanzar una misma estrategia de búsqueda a diferentes buscadores específicos en el campo del Derecho. En este caso, su mayor inconveniente reside en el hecho de estar circunscrito exclusivamente a la búsqueda de información jurídica de sitios norteamericanos.

• PORTALES Y BUSCADORES EN ESPAÑOL DE INFORMACIÓN JURÍDICA

- **VLex** (http://v2.vlex.com/): Uno de los portales jurídicos en nuestro idioma más relevante. Se trata del antiguo portal *Derecho.org* puesto en marcha en 1997 y que tanto prestigio y seguimiento ha tenido entre la comunidad de habla hispana años atrás, reformulado en un servicio más potente y con nuevos contenidos gracias a las inversiones realizadas por el consorcio europeo de capital de riesgo *Grupo 3i.* Además de los servicios característicos de cualquier portal web, ofrece un gran número de recursos jurídicos de todo el mundo accesibles a través de la Web, pudiendo ser consultado su catálogo a través del motor de búsqueda por palabras clave o de categorías temáticas como las siguientes: Civil, Laboral, Nuevas Tecnologías, Mercantil, Público, Penal y Abogados. Asimismo, se encuentran relacionados otros temas de interés jurídico como son el acceso a información legislativa y jurisprudencial en

diversas áreas. En algunos casos el acceso a sus bases de datos exige una suscripción comercial.

Figura I.13: VLex. Portal jurídico en español en Internet.

- **Noticias Jurídicas** (http://noticias.juridicas.com/): otro de los clásicos en nuestro país, creado por la editorial Bosch. Se trata de un importante servicio web de información jurídica en donde es posible acceder de forma gratuita a una base de datos con más de 8.000 documentos a texto completo, con anexos y gráficos, y relacionados hipertextualmente. Además de este importante servicio, se ofrecen otros de gran interés para los juristas y el ciudadano en general, como son: una nutrida selección de artículos doctrinales de prestigiosos juristas, recopilación de software jurídico, informes jurídicos, boletines legislativos y temáticos, etc. Este servicio ha seleccionado más de 4.000 enlaces de interés jurídico existentes en la Web organizados según diversos criterios de acceso: por materias, por su adscripción geográfica o por una clasificación institucional.

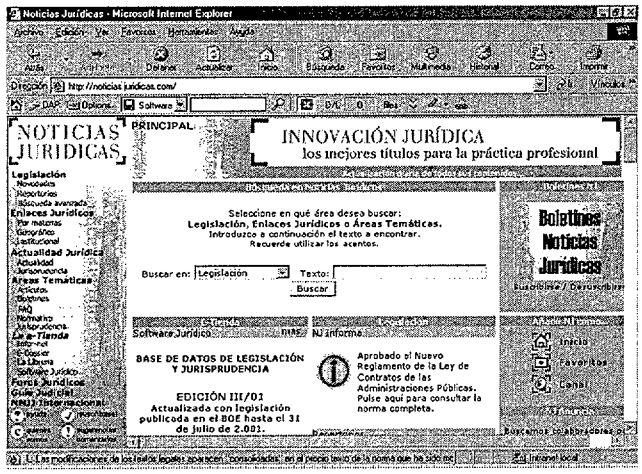


Figura I.14: Noticias.Juridicas.com. Portal jurídico en español en Internet.

- **TuGuiaLegal.com** (http://tuguialegal.metropoliglobal.com/): Otro característico portal web de información jurídica desarrollado en nuestro país. Los contenidos de este sitio se organizan en torno a cinco grandes secciones temáticas, dedicadas a la vivienda y las comunidades de propietarios, la empresa y las sociedades mercantiles, los derechos de los consumidores y el usuario, el derecho civil español y el derecho de las nuevas tecnologías. Este sitio cuenta también con un apartado de formularios jurídicos, un directorio de páginas web de instituciones jurídicas en nuestro país, así como una selección de aquellos recursos considerados como los más útiles para los ciudadanos por su carácter práctico y por ofrecer información gratuita.

- **InfoJuridico.com** (http://www.infojuridico.com/): Otro típico portal web especializado en temas jurídicos orientado claramente al ciudadano común. Además de las noticias jurídicas de mayor actualidad, este servicio nos ofrece una selección de recursos de información en la Web sobre diversas cuestiones y materias relacionadas con el campo del Derecho. Así, las grandes áreas temáticas que este portal contempla son, entre otras, Asesoría Laboral, Boletines Oficiales, Convenios, Estudiantes, Formularios, Guía Judicial, Legislación, Profesionales, Trabajos, así como una selección de enlaces a recursos informativos jurídicos considerados de interés.

- **IurisLex** (http://www.iurislex.net/): Se especializa principalmente en la recopilación de noticias de interés jurídico en nuestro país procedentes de cualquier ámbito del Derecho. Dispone de un potente motor de búsqueda para la localización de información electrónica, según diversos criterios (acotación por el tipo de documento, por la materia, por una determinada categoría, así como la utilización de los operadores booleanos Y y O).

- **CanalJuridico.com** (http://www.canaljuridico.com/): Uno de los últimos portales web españoles de contenido jurídico aparecidos en Internet pero, según sus patrocinadores, con la vocación de constituirse en el primer portal jurídico en español existente en Internet. Sigue los planteamientos habituales para este tipo de servicios, con la inclusión de noticias jurídicas de actualidad, pero destaca especialmente por la inclusión de diversas bases de datos jurídicas de acceso gratuito, entre las que se encuentran lo publicado en el BOE, arquitectura jurídica, formularios, Jurisprudencia de diversos Tribunales españoles, Legislación, y otras menores sobre noticias jurídicas, apuntes y fondo editorial. Su base de datos de jurisprudencia es ciertamente exigua y muy limitada en prestaciones, tanto de recuperación de información como de visualización de la misma (los documentos no están relacionados hipertextualmente).

Figura I.15: CanalJurídico.com. Portal jurídico en español en Internet.

- **InfoDerecho** (http://www.infoderecho.com/): Estructurado como un clásico directorio de recursos en la web, divide los enlaces seleccionados en 14 áreas temáticas: Abogados y Asociaciones, Administración, Base de datos legislativa, Derechos, Economía y Negocios, Internacional, Editoriales Jurídicas, Jurisprudencia, Legislación, Registro Mercantil y de la Propiedad, Medios de Comunicación, Otras Profesiones, Universidad, y Páginas Personales.

- **TodaLaLey.com** (http://www.todalaley.com/): Se trata un servicio de búsqueda de información jurídica especializado en la localización de textos legislativos aparecidos en los diversos Boletines Oficiales de nuestro país. Además del buscador por palabras clave para la localización de dichos textos, agrupa de forma temática diferentes modelos de contratos (civiles y mercantiles) y Formularios (administrativos y procesales). Contiene un apartado dedicado exclusivamente a todo lo relacionado con la Legislación en nuestro país (últimas Leyes publicadas dentro de diversas categorías, como Leyes Orgánicas, Leyes Ordinarias, Reales Decretos Leyes, Reales Decretos Legislativos, Leyes de las distintas Comunidades Autónomas, etc.). Igualmente se recogen enlaces a recursos informativos sobre diversas cuestiones relacionadas con la Administración Pública, así como un servicio de guías y manuales electrónicos relacionadas con el Derecho Civil, Mercantil, Laboral, Administrativo y Procesal.

• BASES DE DATOS CON INFORMACIÓN JURISPRUDENCIAL

- **BRISA: Base de datos Relacional de la Industria y Servicios Ambientales** (http://www.mcyt.es/brisa/): Creada en un principio por el Ministerio de Industria y Energía español en colaboración con la Asociación Nacional de Fabricantes de Bienes de Equipos. Se trata de una base de datos con información legislativa y jurisprudencial emitida por diversos Tribunales de nuestro país en materias como las energía

alternativa, equipos y montajes, gestores de residuos, ingeniería y servicios, así como información sobre patentes españolas.

- **CELEX** (http://europa.eu.int/celex/): se trata de la gigantesca base de datos jurídica de la Unión Europea elaborada por la Oficina de Publicaciones Oficiales de las Comunidades Europeas, y de la cual iremos hablando en algunos otros apartados de esta tesis doctoral. Es la base de datos del Derecho Comunitario por excelencia, disponible en todas las lenguas oficiales de la UE. En ella es posible encontrar actos jurídicos, tratados, legislación, dictámenes y resoluciones de las distintas instituciones de la UE, así como la jurisprudencia del Tribunal de Justicia. Se trata de una base de datos de acceso restringido (acceso comercial), donde se ofrecen diversas formas de interrogación: el *Menu searching* es la búsqueda asistida a través de menús, y la *Expert searching*, opción más potente y sofisticada que exige un conocimiento perfecto de los distintos comandos y opciones de interrogación. También es posible acceder al *Celex basic service*, servicio básico y acceso gratuito a Celex. Es el que utilizan los funcionarios de la UE desde hace varios años.

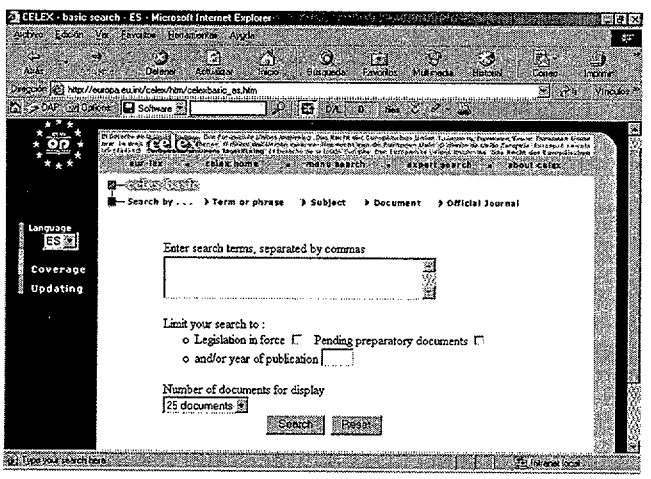


Figura I.16: CELEX. Base de datos jurídica en Internet de la Unión Europea.

- **ICAL Cerca de sentencies** (http://obertacat:co.set@www.juridica.com/ical/bstas.htm): Se trata de la base de datos de sentencias y resoluciones de la Audiencia Provincial de Lleida creada por el *Il.lustre Col.legi d'Advocats de Lleida* (ICAL) en 1990. Su acceso está restringido.

- **TRIBUNAL DE JUSTICIA DE LA UE** (http://curia .eu.int/es/jurisp/index.htm): el Tribunal de Justicia de las Comunidades Europeas dispone de una bases de datos jurídica propia en la cual se recoge la jurisprudencia reciente del Tribunal de Justicia y del Tribunal de Primera Instancia de la UE. Su acceso es gratuito y las sentencias se encuentran a texto completo.

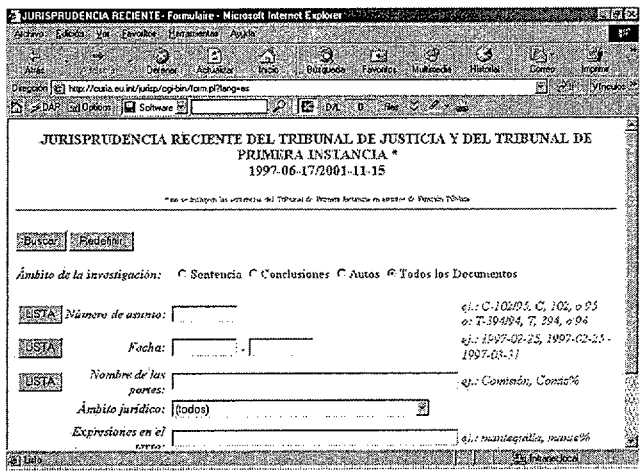


Figura I.17: Base de datos jurídica en Internet del Tribunal de Justicia de la Unión Europea.

- **CDE DE LA UNIVERSIDAD DE ALICANTE** (http://fcae.ua.es/cde/lex.htm): El Centro de Documentación Europea de la Universidad de Alicante ha creado su propia base de datos de legislación y jurisprudencia de la Unión Europea, en donde la búsqueda del documento se realiza a través de la referencia al mismo (Reglamento, Directiva, decisión, Documento COM o Asuntos del Tribunal de Justicia).

- **TRIBLEX** (http://ilis.ilo.org/ilis/trib/ilseartr.htm#SearchFormE): Base de datos jurídica creada por la Organización Internacional del Trabajo (OIT) dentro del *ILO/ILIS Referral System*. Se trata de una base de datos jurisprudencial de acceso público en donde se contienen las decisiones de jurisprudencia emitidas por el Tribunal Administrativo de esta organización internacional. Se encuentra disponible tanto en francés como en inglés y, como ya se comentó con anterioridad, dispone de un tesauro jurídico de ayuda para la consulta correcta de términos.

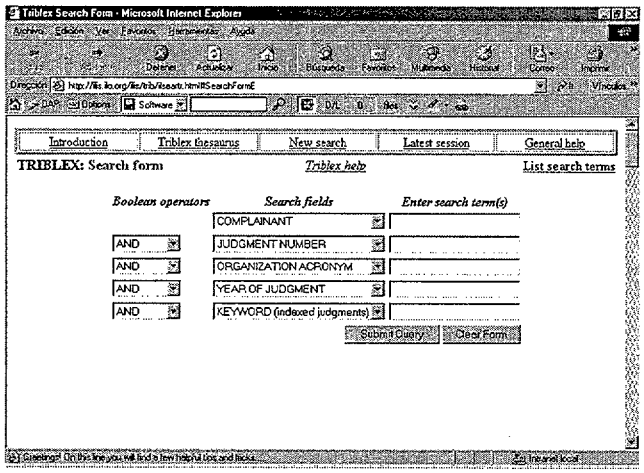


Figura I.18: TRIBLEX. Base de datos jurídica en Internet del Tribunal Administrativo de la OIT.

El panorama descrito es, por supuesto, muy cambiante. No sólo porque no dejan de aparecer en escena nuevos proveedores de contenido sino porque además éstos incorporan continuamente nuevos servicios a los ya ofrecidos. En cualquier caso, podemos afirmar que en los últimos años el potencial de Internet y la Web ha sido descubierto y está comenzando a ser explotado por las empresas y organismos públicos que tienen intereses en la difusión de información jurídica, como se verá y analizará en la próxima Parte de esta tesis.

CONCLUSIONES A LA PARTE

A lo largo de esta primera Parte se ha venido analizando el importante papel que juega en toda sociedad moderna el correcto tratamiento y acceso a la información jurídica producida por las diferentes instituciones y organismos de la Administración de Justicia. Para poder llevar a cabo esta labor fundamental la ciencia del Derecho se ha venido apoyando en una serie de técnicas y procedimientos definidas por otras ciencias, en especial las derivadas de la ciencia de la Documentación y de la ciencia de la Computación. Esta combinación de saberes ha dado lugar a nuevas disciplinas científicas, como es el caso de la informática jurídica documental en donde su objetivo final consiste en la confección de bases de datos jurídicas (legislativas, jurisprudenciales y/o doctrinales) de interés tanto para los profesionales del Derecho como para el ciudadano en general. De igual modo se ha venido analizando algunos aspectos de interés relacionados con las aplicaciones informáticas para el tratamiento de la información judicial y, de forma más concreta, a la jurisprudencia emanada de diversos Tribunales de Justicia de nuestro país, en donde el Tribunal Constitucional y la jurisprudencia emanada de él juegan un papel determinante en el ordenamiento democrático y constitucional de nuestro país. Finalmente, se ha analizado igualmente la irrupción de la red Internet en las sociedades actuales y cómo este fenómeno tecnológico y social ha provocado grandes cambios en las formas actuales de difusión y acceso a la información jurídica en todo el mundo.

A modo de resumen, y como conclusiones a esta tercera parte de esta tesis doctoral se pueden establecer los siguientes puntos de interés:

- El Derecho desde sus orígenes ha venido manteniendo una necesidad imperiosa de ser conocido por los ciudadanos a los que afectan sus normas y regulaciones, siendo plasmadas éstas en un soporte documental para su correcta transmisión; los denominados *documentos jurídicos*. Estos documentos tienen unas peculiaridades que les hacen diferenciarse en gran medida del resto de documentos producidos por otros sectores de la Sociedad, como son las de poseer validez por sí mismos pero difícilmente entendibles fuera del conjunto que forma el ordenamiento jurídico de un país o región;

así como la definición de unas estructuras claras y consolidadas por la tradición jurídica, con un lenguaje científico sumamente característico, difícilmente entendible por aquellas personas ajenas a esta ciencia.

- La Documentación ha proporcionado los mecanismos necesarios para tratar y analizar adecuadamente la información contenida en los documentos jurídicos, abstrayendo de ellos las partes más sustanciales para una correcta descripción de los mismos así como normalizando los conceptos principales contenidos en el texto a través de la transformación del lenguaje jurídico, ambiguo y con algunas vaguedades en su terminología, a un lenguaje documental, más preciso y certero. Fruto de esta estrecha relación entre el Derecho y la Documentación a lo largo del sigo XX surge una nueva disciplina científica conocida por el nombre de Documentación jurídica, la cual pretende dar solución a los problemas derivados de la constante producción de documentos jurídicos en las sociedades modernas a través de una serie de técnicas y procedimientos fuertemente consolidados. De este modo, las técnicas procedentes de la Documentación permiten generar desde los textos jurídicos originales una serie de productos secundarios de vital importancia para la localización y recuperación de información y datos jurídicos ante la demanda de una necesidad puntual de información por parte del jurista.

- La ciencia de la computación aplicada al Derecho tiene, de igual forma, una importancia capital en toda esta labor de control y acceso a la información jurídica. La jurimetría y la iuscibernética han venido ha potenciar, en un primer estadio de desarrollo, ese papel que jugaba la Documentación en cuanto a los procesos de tratamiento y recuperación de la información jurídica, acelerando la realización de dichos procesos y aumentando hasta cotas muy elevadas las posibilidades de almacenamiento, gestión y recuperación rápida y precisa de este tipo de información. De esta unión entre el Derecho y la informática han surgido dos disciplinas fundamentales en nuestros días: El Derecho informático, el cual estudia y analizar las implicaciones jurídicas que las nuevas tecnologías tienen en nuestra sociedad y trata de ordenar jurídicamente dichas relaciones, y la Informática jurídica, cuyo principal

objetivo radica en la creación de sistemas informáticos de bases de datos que sean capaces de almacenar y recuperar de forma adecuada la información jurídica. Por extensión de la idea de la aplicación de la Informática al Derecho, la informática jurídica se ha venido dividiendo en diferentes ramas según la aplicación que se diera de la misma: informática jurídica de gestión, informática registral, informática notarial, informática operacional, informática decisional e informática jurídica de ayuda a la decisión, en donde se enmarca la actual informática jurídica documental, de aplicación a los diversos documentos jurídicos (legislación, jurisprudencia y doctrina) para la creación de sistemas informáticos capaces tratar de forma adecuada o con criterios documentales todo este vasto conjunto de información.

- El desarrollo más notorio de la aplicación de las ciencias de la Documentación y la Informática al Derecho se encuentra, sin duda, en las bases de datos jurídicas. La necesidad existente en las sociedades actuales de tener un conocimiento preciso y exacto del ordenamiento jurídico, en la disponibilidad y accesibilidad a las fuentes del Derecho, en la Exhaustividad en la recuperación de información legislativa y jurisprudencial, así como en otras cuestiones que garanticen la seguridad y veracidad de la información jurídica hacen que este tipo de dispositivos electrónicos de almacenamiento y recuperación de información hayan venido desarrollándose e incrementándose de forma tan espectacular a lo largo de su historia. Las técnicas y los procedimientos para el almacenamiento y la recuperación de la información jurídica contenida en estos sistemas electrónicos han ido variando a lo largo del tiempo con los nuevos desarrollos tecnológicos aparecidos pero en todos los casos las técnicas de descripción y análisis documental de la información contenida han constituido la base del desarrollo de estas bases de datos. Las bases de datos jurídicas españolas, en donde la tradición en su comercialización ha sido la nota dominante, se han ido adaptando de igual modo a estos nuevos avances tecnológicos, desde sus primeras versiones en soporte CD-ROM hasta los actuales sistemas en DVD y acceso a través de Internet, con nuevos y espectaculares desarrollos técnicos que facilitan enormemente su interrogación y la consulta de la información jurídica contenida.

- El desarrollo de la informática jurídica en nuestro país, y en especial la informática aplicada a la Administración de Justicia, ha sido ciertamente tardía en comparación al resto de países de nuestra esfera socio-política. De forma general, se puede decir que este desarrollo ha venido ligado a las iniciativas que de forma particular y con gran desconexión entre ellas se han venido realizando en diferentes partes del territorio nacional y por distintos órganos administrativos y judiciales de la compleja maraña jurisdiccional de nuestro país. Muchas han sido las iniciativas encaminadas a crear centros de control y normalización de las técnicas informáticas que se deberían emplear en los diferentes Tribunales y Juzgados de nuestro país pero en la mayoría de estos casos dichas iniciativas han venido fracasando estrepitosamente. Sólo hasta fechas recientes se ha conseguido un cierto nivel de consenso en este asunto, consenso promovido desde los propios órganos de la Administración del Estado (caso de las iniciativas llevadas a cabo desde el Ministerio de Justicia) o de los órganos superiores de la Administración de Justicia (caso de las iniciativas propuestas por el Consejo General del Poder Judicial). En la actualidad, y con la puesta en macha de proyectos de coordinación y centralización documental, como es el caso de la constitución del Centro de Documentación Judicial (CENDOJ) por parte del CGPJ, de proyectos de informatización de las oficinas judiciales, caso del programa LIBRA, y del consenso alcanzado entre los principales partidos políticos de nuestro país para la reforma de la Justicia, es posible hablar ya de un panorama bastante más esperanzador en todos estos asuntos.

- El tratamiento documental e informático de la información jurisprudencia emanada por los diversos Tribunales de Justicia de nuestro país ha sufrido igualmente los avatares de estas transformaciones políticas, sociales y tecnológicas que se han venido sucediendo a lo largo de estas últimas décadas. En el caso de la jurisprudencia emanada del Tribunal Constitucional toma una especial importancia dado que ésta tiene además un valor normativo al poder modificar o derogar otras normas jurídicas. Las sentencias del Tribunal Constitucional juegan, por tanto, un papel fundamental en nuestra sociedad de

Derecho y su difusión y accesibilidad por parte de todos los ciudadanos del Estado español debe ser garantizado por los distintos poderes públicos y judiciales del país.

- Hasta fechas recientes, el tratamiento y difusión electrónica de la información jurisprudencial era realizado por diversas instituciones editoriales, públicas o privadas, especializadas en el campo del Derecho y en soporte CD-ROM. Su accesibilidad estaba, por tanto, limitada a los profesionales de la actividad jurídica y a aquellos otros profesionales, investigadores o estudiosos del Derecho que podían acceder a centros de difusión y consulta de estas bases de datos (centros de información jurídica, bibliotecas especializadas, etc.). El tratamiento a que es sometida la información jurisprudencial en estas bases de datos difiere bastante poco de un sistema a otro, aunque las capacidades del software de recuperación así como la presentación de los documentos al usuario, han venido marcando las verdaderas distinciones entre estos productos. La irrupción de las tecnologías derivadas de la red Internet ha supuesto un nuevo revulsivo para estas empresas dado que ahora la difusión que de los datos jurídicos se puede hacer es mucho mayor y de forma más sencilla. En la actualidad, y para poder justificar su labor, estas empresas deben ofrecer un importante valor añadido a la información jurídica primaria, así como a los mecanismos para su difusión, localización y recuperación precisa y exacta de la información demandada pues, como veremos en el siguiente punto, han entrado en escena multitud de servicios en Internet que vienen a competir con estos servicios comerciales, ofreciendo de un modo gratuito el acceso a grandes volúmenes de información jurídica de todo tipo.

- La aparición de la red Internet en nuestra sociedad y la posibilidad que se tiene para establecer una difusión de la información producida por las diferentes Administraciones del Estado ha venido a revolucionar en gran medida el acceso a la información jurídica por parte del ciudadano. En primer lugar, los distintos gobiernos democráticos del planeta han considerado a la red Internet como una herramienta ideal para acercar la Administración Pública al ciudadano. De otra parte, este acercamiento ha producido una mayor exigencia por parte del ciudadano en cuanto a la disponibilidad libre y

gratuita de la información que es generada desde los diversos Poderes del Estado. En este contexto se ha venido esgrimiendo que si la información jurídica, especialmente la legislativa y la jurisprudencial, las cuales han de ser de obligado conocimiento por parte del ciudadano y son emanadas desde las correspondientes instituciones y organismos públicos que son sustentadas económicamente por los ciudadanos resulta, por tanto, completamente justificable que dicha información esté a disposición de los mismos a través de los mecanismos que aporta la red Internet para la difusión masiva de información. Al amparo de esta idea, han venido surgiendo en diferente países del mundo, en donde España no constituye una excepción, diversos servicios de información jurídica en Internet, poniendo de este modo a disposición de los ciudadanos a través de la red Internet la información legislativa y jurisprudencial que les puede afectar. Este tipo de servicios jurídicos en la WWWW, ha sufrido un fenómeno de "portalización" por lo que en la actualidad estos portales se configuran como auténticos centros globales de información jurídica orientados a satisfacer las necesidades informativas de los profesionales del Derecho, investigadores y estudiosos del mismo y, especialmente, del común de los ciudadanos.

- Las nuevas bases de datos jurídicas que se han venido construyendo en estos últimos años al amparo de las tecnologías de la WWW, en especial a través de su lenguaje más característico, el HTML, han revolucionado en gran medida el concepto tradicional de acceso a la información contenida en los antiguos modelos basados en los soportes ópticos. En este caso se trata de grandes conjuntos de documentos a texto completo marcados con las etiquetas establecidas por dicho lenguaje y en donde las relaciones explícitas que se dan en sus textos son puestas de manifiesto a través de la inclusión del correspondiente ancla hipertextual. Ello permite ir navegando a través de un documento a otro según los intereses existentes para cada consulta de información por parte de los usuarios. La inclusión de un motor de búsqueda permite optar por la localización de una o varias palabras dentro del texto completo de los documentos que conforman la base de datos hipertextual. Este último aspecto es el que de forma más seria ha venido limitando el desarrollo de este tipo de bases de datos dado que el

lenguaje HTML, como se explicará detenidamente en capítulos posteriores, no proporciona la suficiente capacidad para una descripción semántica completa de los distintos elementos informativos que se integran dentro de dichos textos. La solución a este grave problema vendrá de la mano de otro de las tecnologías surgidas en el entorno de la WWW: el metalenguaje de marcado de texto XML.

En la siguiente Parte de esta Tesis doctoral se expondrán las diversas técnicas y métodos que han venido surgiendo para el tratamiento de los documentos electrónicos, en donde el desarrollo de diversos lenguajes de marcado de los textos contenidos, tanto para su presentación como para su descripción estructural y/o semántica, constituirá el eje central de la exposición de los siguientes capítulos.

PARTE II

LOS LENGUAJES DE MARCADO DE DOCUMENTOS ELECTRÓNICOS

INTRODUCCIÓN A LA PARTE

No resulta ajeno en nuestros días, y por parte de todo tipo de profesionales, hablar acerca de y hacer uso de tecnologías informáticas tan populares como son las asociadas al fenómeno Internet y los servicios a los que da acceso. En especial, el servicio más característico de la red Internet, la *World Wide Web* (WWW), es hoy día utilizado por un gran número de personas y organizaciones de todo el mundo, ya sea para fines profesionales, comerciales o, simplemente, como un elemento más incorporado al ocio y diversión en el espacio doméstico[1]. Para muchas de estas personas lo que ya no es tan conocido es todo lo que existe *por debajo* de esa gran maraña de información electrónica: aspectos tales como los tipos de redes de transporte utilizados, velocidades de transmisión, protocolos de comunicaciones, modelo cliente-servidor, lenguajes y formatos de creación de documentos electrónicos, etc., son términos y conceptos desconocidos para muchos de estos usuarios no expertos desde el punto de vista tecnológico. En el caso concreto del modelo que sustenta y da soporte a los documentos electrónicos que llenan de contenido el macroespacio virtual que constituye la WWW, el lenguaje HTML (*HyperText Markup Language*), dicho desconocimiento es, en muchos casos, extremo. El usuario doméstico sólo demanda de la Red que le sean servidos unos determinados contenidos informativos que son materia de su interés, en general, todo lo relacionado con el ocio, entretenimiento y diversión, siendo, por tanto, sus únicas preocupaciones aspectos tales como la búsqueda y localización de dicha información, la rapidez en la transferencia, la calidad de los

[1] Resulta difícil obtener datos estadísticos exactos y fiables relativos al tamaño de la Web y al número de usuarios a escala mundial conectados a la red Internet. La *Internet Society* (http://www.isoc.org/) recoge enlaces a una serie de instituciones y organismos que se dedican a analizar estos temas. Así, por ejemplo, la compañía irlandesa NUA realiza periódicamente un sondeo para calcular el número aproximado de usuarios conectados a esta gran red de ordenadores, estableciendo a fecha de verano de 2000 una cifra de 332,73 millones de usuarios en todo el mundo (http://www.nua.ie/surveys/how_many_online/index.html). La fuente de datos más fiable para conocer y analizar esta situación en España lo constituye el Estudio General de medios que realiza la Asociación para la Investigación de Medios de Comunicación (http://www.aimc.es/), pues en él se resumen las estadísticas que se van publicando en diferentes medios de comunicación.

contenidos y los gastos económicos ocasionados en todo este proceso de conexión telemática. Otros usuarios tecnológicamente más curiosos deducen, descubren y llegan a aprender el lenguaje de construcción de documentos HTML, o *páginas Web*, como popularmente se los conoce, incorporándose de este modo a la gran comunidad de redactores que constantemente nutren de información el espacio Web[2]. Pero, en líneas generales, son muy pocas las personas que realmente comprenden el proceso de análisis y estructuración de la información que están llevando a cabo cuando construyen documentos electrónicos marcados con el lenguaje HTML, haciendo extensible esta afirmación a un gran número de profesionales de múltiples áreas de actividad que realizan de forma más o menos esporádica esta labor documental.

El lenguaje HTML no es más que un lenguaje de marcado de texto electrónico (en su expresión original, un *mark-up language*), mediante el cual se identifican ciertos elementos que suelen aparecer comúnmente en cualquier texto de propósito general (un título, un cuerpo de documento que puede contener párrafos, tablas de datos, listados, imágenes, etc.) y se los encapsula o *etiqueta* para que sean posteriormente reconocidos e interpretados por diversas aplicaciones informáticas, especialmente por visualizadores de documentos HTML, comúnmente conocidos por el nombre de *navegadores*.

Pero para los efectos de esta tesis, lo que realmente nos interesa y preocupa es conocer y entender el proceso de génesis y evolución de los lenguajes de marcado de texto en documentos electrónicos hasta llegar al modelo más actual y pujante, y aquí defendido, el metalenguaje XML (*Extensible Markup Language*), de aplicación no sólo en el variopinto espacio Web, sino también en los microespacios virtuales que constituyen los documentos electrónicos de una determinada institución o, también denominados, de propósito específico. Así pues, se hace necesario ir a la raíz de todo ello, a los grandes pilares de todo este proceso evolutivo, constituido, por un lado, por la importancia de la marca y los lenguajes de marcado de texto para definir los elementos estructurales y de contenido de un

[2] La organización *The Censorware Project* recoge varios estudios aparecidos en Internet y en diversas revistas científicas, como *Science* y *Nature*, llegando a establecer el contenido de la Web en 2.360 millones de páginas, incrementándose diariamente dicha cifra con casi 5 millones de páginas nuevas. Para una mayor información

determinado modelo de documento electrónico, y, por el otro, la búsqueda de un formato sencillo y universal de documento electrónico que haga factible el paradigma de efectividad en la gestión de información (creación, almacenamiento, búsqueda y recuperación) en entornos informatizados.

véase *Size of the Web* [documento HTML]. The Censorware Project, 26 de enero de 2000. Disponible en http://censorware.org/web_size/ (consultado el 14 de agosto de 2000).

CAPÍTULO II.1

INTRODUCCIÓN A LOS LENGUAJES DE MARCADO DE TEXTOS ELECTRÓNICOS

II.1.1. ORIGEN Y CONCEPTO DE MARCA Y LENGUAJE DE MARCADO DE TEXTO: DEL DOCUMENTO IMPRESO TRADICIONAL AL DOCUMENTO ELECTRÓNICO

La palabra **marca** puede ser definida e interpretada de múltiples maneras, como así lo pone de manifiesto la Real Academia de Lengua en su Diccionario de la Lengua Española al contemplar las múltiples definiciones existentes para dicha voz en nuestro idioma. Entre todas ellas, y para nuestros propósitos, recogemos las definiciones 6, como la "Acción de marcar", y la 7, como la "Señal hecha en una persona, animal o cosa, para distinguirla de otra, o denotar calidad o pertenencia"[3]. La acción de establecer marcas, esto es, **marcar**, es definida igualmente en este diccionario como "Señalar con signos distintivos". Igualmente ilustrativa resulta la definición de "marcar" contemplada en el diccionario de María Moliner, al indicar que se trata de "Poner una marca en una cosa para distinguirla o para hacerla notar"[4]. El término **marcado** es contemplado en nuestro idioma, además de cómo participio pasivo del verbo *marcar* como un adjetivo utilizado con el significado de hacer muy perceptible algo (un acento o variante lingüística, o una actitud personal ante un hecho), acepción éste que se aleja del propósito con el que se utilizará en esta tesis doctoral. Por tanto, la expresión *lenguaje de marcado* con el significado de *lenguaje de marca* es incorrecta. Así, aunque lingüísticamente sería más correcto hablar de *la marca de algo* (pues es la acción de marcar), la tradición profesional en la esfera de la documentación científica ha preferido, como se verá posteriormente, utilizar el participio pasado para denominar la acción que se está llevando a cabo, al ser mucho más claro y explícito cuando es aplicado a los

[3] Real Academia Española. *Diccionario de la lengua española* (21ª ed.). Madrid: RAE, 1992, p. 935.

[4] María Moliner. *Diccionario de uso del español* (2ª ed.) Madrid: Gredos, 1998, Tomo II, p. 277.

documentos (véase, el matiz semántico diferencial que existe entre las expresiones "la marca de un documento" y "el marcado de un documento").

Antes de analizar y establecer el consiguiente discurso narrativo sobre el alcance de dichas acepciones resulta conveniente transcribir aquí las definiciones correspondientes a las anteriores palabras para la lengua inglesa, dado que a lo largo de este apartado recurriremos asiduamente a dichas voces en este idioma. Así, el diccionario de la Universidad de Oxford, define a la palabra *mark* en su segunda entrada como *a line, figure, or symbol made as an indicator or record for something,* y a la acción de marcar como *write a word or symbol on (an object), typically for identification*[5].

Ambas lenguas, por tanto, coinciden significativamente en la interpretación de las palabras marca y su correspondiente acción: la marca constituye un signo o señal realizada con cualquier instrumento y sobre cualquier medio que se establece para destacar o diferenciar frente al resto de cosas aquello sobre lo cual se ha ejercido el acto de marcar. Todo ello, la marca y su acción, es algo incorporado plenamente a nuestra vida y en todos los órdenes de la misma; es una de las acciones que el ser humano realiza constantemente. Baste simplemente echar un vistazo a nuestro alrededor para comprobar que estamos rodeados de marcas (signos o señales) que hacen que se establezcan diferencias entre unos objetos y otros, o, dentro de un mismo grupo, entre los elementos integrantes de una misma familia[6].

[5] Judy Pearsall (ed.) *The New Oxford Dictionary of English.* New York [etc.]: Oxford University Press, 1998, p. 1132.

[6] Haciendo uso de un ejemplo sencillo de la vida cotidiana, podemos contemplar y distinguir distintas marcas de coches, reconocidas por el signo o señales visibles que los fabricantes insertan en la carrocería de los mismos (normalmente, el logotipo de la empresa). Dentro de un mismo fabricante, los diferentes modelos de automóviles que se lanzan al mercado son distinguidos con diferentes nombres y versiones, igualmente explícitos en el propio auto. Incluso en el caso de coches completamente idénticos en sus marcas, modelos y características, siempre hay algún elemento que lo hace único y distinguible de los demás (número de bastidor, número de matricula, artículos decorativos incluidos por los propietarios de cada vehículo, etc.)

Restringiendo el campo de aplicación de estos conceptos a la materia de estudio y análisis de la disciplina científica en la cual se inscribe esta tesis doctoral, la Documentación científica, y al objeto de su estudio, los documentos, es fácil comprobar que lo anteriormente expuesto se cumple de igual modo. Así, aplicado a los documentos impresos, y al nivel más extremo, algunos autores señalan que todos éstos llevan en esencia un conjunto de signos ordenados mediante unas reglas de inserción que les hace ser comprensibles para el ser humano. Cuando un autor escribe algo, está realizando un ejercicio de marcado (por ejemplo, la asignación de espacios en blanco indica una separación entre las palabras, las comas y los puntos indican espacios de separación entre frases, signos o marcas de pregunta, admiración, puntos suspensivos, etc.). Por otro lado, también se puede establecer que unos documentos se distinguen de otros mediante elementos que señalan su autoría, el título concreto de la obra, y, en el caso de muchos documentos impresos, otros datos relativos a la edición e impresión.

Pero desde un punto de vista más restrictivo, y para los efectos de esta tesis, el concepto de la marca es originario del mundo de la edición y tiene, pues, una ya larga historia. Antes incluso de la invención de la imprenta, muchos manuscritos y códices, confinada su producción principalmente en manos de escritorios monásticos o capitulares, eran anotados con una serie de signos o señales marginales que establecían las pautas que debían seguir los copistas a la hora de enfrentarse a la duplicación de dicha obra o, en otros casos, corregían el trabajo ya realizado por éstos. Pero es con la llegada de la reproducción mecánica de los documentos mediante la imprenta cuando la utilización y difusión de la marca se hacen más generales. En el ámbito de la publicación impresa tradicional, y hasta fechas muy recientes, los manuscritos originales y las pruebas de imprenta eran anotadas con sencillas instrucciones para el correcto procesamiento por parte del compositor o cajista de la imprenta. Este acto de anotar y preparar documentos, es decir, la *corrección*, es definida en el ámbito de la impresión clásica por Martínez de Sousa como la operación que consiste en suprimir las equivocaciones que contengan el original o las pruebas, por lo cual

se denomina *corrección de estilo* la primera y *corrección tipográfica* la segunda"[7]. Este doble trabajo de corrección, tanto del fondo como de la forma, tiene su manifestación en la utilización de una serie de marcas manuscritas que el corrector va insertando tanto en aquellas partes del documento que son dignas de destacar (error detectado) como en uno de los márgenes del mismo (corrección del mismo), quedando de este modo señaladas para su posterior procesamiento. Este tipo de marcas o instrucciones de procesamiento es conocido en el ámbito de la edición e imprenta tradicional por el nombre de *signos de corrección* [8].

Tomando el sentido más amplio de la palabra **lenguaje**, en el que todo conjunto de señales que dan a entender algo es entendido como tal[9], podemos establecer que el conjunto formado por todos estos signos de corrección debe ser, por tanto, contemplado como un auténtico lenguaje, de uso e interpretación por la comunidad de profesionales pertenecientes al ámbito de la edición impresa tradicional. Si en un lenguaje toda señal tiene un significado, debe existir un mecanismo explicativo que identifique inequívocamente significante y significado para, de este modo, ser interpretado de forma correcta y adecuada por todos los individuos que hagan o vayan a hacer uso de él. En el contexto de la edición impresa tradicional este mecanismo se plasma en el llamado *cuadro de signos de corrección*. El primer cuadro de signos de corrección del que se tiene constancia data de 1773 y se le atribuye al francés Pierre-François Didot, codificando y unificando todos los signos que convencionalmente se venían utilizando en los talleres de impresión de aquella época y país[10].

---

[7] José Martínez de Sousa. *Diccionario de tipografía y del libro* (2ª ed.) Madrid: Paraninfo, 1981, p. 56.

[8] Aunque no es propósito de esta tesis profundizar en el tema de los usos y técnicas en las correcciones de imprenta, sí nos interesa señalar brevemente aquí algunas características propias de esta actividad dada la importancia que ello conlleva para comprender adecuadamente la utilización de marcas y lenguajes de marcado de texto en documentos electrónicos.

[9] Real Academia de la Lengua. *Op. cit.*, p. 878.

[10] John Dreyfus, François Richaudeau (dir.) *Diccionario de la edición y de las artes gráficas.* Salamanca [etc.]: Fundación Germán Sánchez Ruipérez, 1990, p. 645.

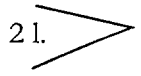
Analizando muy brevemente, y de forma muy general, el contenido de todo cuadro de signos de corrección, se pueden establecer dos tipos bien diferenciados de signos[11]:

1. Aquellos cuya función es simplemente la de señalar aquella parte del texto (letra, palabra, frase o lugar) que haya de ser corregida, conocido por el nombre de *llamadas.*

   Algunos ejemplos de este tipo de signo son:

2. Aquellos cuya función es la de indicar la operación que se ha de llevar a cabo en el lugar señalado por el corrector a través de la llamada, agrupándose aquí las llamadas *enmiendas* y *señales.*

   Algunos ejemplos de este tipo de signos son:



Poner la cantidad de blanco indicada.

Unir, quitar espacio.

Evitar dos palabras iguales en líneas consecutivas.

Cursiva.

Por último, y por los futuros paralelismos que se establecerán en esta tesis, es importante destacar algunas de las características esenciales del lenguaje generado por los signos de corrección que se han venido utilizando tradicionalmente en el mundo de la edición impresa, como son su internacionalidad (se usan en la práctica totalidad de los países, con ligeras diferencias) y la ausencia de ambigüedad en el significado de cada uno de

[11] José Martínez de Sousa. *Op. cit.*, p. 252.

los signos (cada signo tiene un significado único convencional, entendible por los profesionales de ese campo de actuación).

En este entorno tradicional de producción de documentos podemos hacer, por tanto, una primera aproximación general al concepto de marcado de texto. El **marcado de texto**, afirmando que sería todo aquello que no forma parte del contenido original del texto pero que se introduce para informar acerca de algo sobre el mismo[12]. De hecho, cuando el texto escrito se verbaliza a través del habla, muchos de los signos de marcado del texto se hacen patentes haciendo uso para ello de gestos y variaciones paralingüísticas para transmitir de forma correcta el verdadero sentido de la información contenida en ese espacio (por ejemplo, los signos de admiración hacen que enfaticemos las palabras contenidas entre éstos, elevando la voz y transmitiendo una determinada información con una carga expresiva concreta).

Con la aparición y aplicación de la informática a los procesos de fotocomposición en las modernas imprentas la utilización de estas técnicas de corrección de originales y pruebas a través signos manuscritos cayó en desuso, hasta su prácticamente completa desaparición. Pero su filosofía subyacente, la idea de marcar el texto de los documentos a través de signos o marcas para su posterior presentación, se extendería a este nuevo entorno informatizado de trabajo con documentos electrónicos. Será aquí donde marca y lenguaje de marcado tomarán su verdadero sentido y mayor difusión. De hecho, en la actualidad, con la incorporación de los sistemas informáticos de procesamiento de texto a nuestra vida cotidiana, tanto personal como profesional, realizamos el acto del marcado del texto redactado sin ser muchas veces conscientes de este hecho. Ya estamos tan acostumbrados a redactar los textos preparándolos para su posterior lectura, en pantalla o en papel impreso, que no somos conscientes, las más de las veces, de que cuando ese texto es almacenado en un fichero electrónico lleva asignado un determinado marcado (o codificación) de

[12] James H. Coombs, Allen H. Renear, Steven J. DeRose. *Markup Systems and the Future of the Scholarly Text Processing* [documento HTML]. OASIS, [sin fecha]. Disponible en http://www.oasis-open.org/cover/coombs.html (consultado el 12 de julio de 2000).

presentación, el cual hace mención al modo de proceder que una determinada aplicación informática debe seguir para interpretar correctamente el texto marcado (negrita, cursiva, tamaños de fuentes, etc.).

II.1.2. LOS LENGUAJES DE MARCADO EN LOS DOCUMENTOS ELECTRÓNICOS

II.1.2.1. INTRODUCCIÓN:

De forma simple, se puede decir que un documento electrónico no es otra cosa que una serie de ceros y unos capaces de ser almacenados y manipulados por ordenadores y suministrados a través de redes informáticas[13]. Pero desde un punto de vista más cercano a la ciencia de la Documentación, el documento electrónico es entendido como la representación de un documento en la forma de una estructura de datos informáticos interpretables por la memoria de un ordenador y transportables a otro[14].

Pero en realidad, el documento electrónico lleva asociadas otras características que le hacen ser una entidad mucho más compleja y que determinan la función y el propósito para el que ha sido generado y, en buena medida, la forma en que puede ser almacenado, recuperado y suministrado. Nos estamos refiriendo aquí a los múltiples formatos de codificación o marcado de documentos electrónicos que han existido y/o existen desde que es posible el procesamiento de éstos por parte de las computadoras. Un formato de documentos electrónicos supone, por tanto, un conjunto de reglas o convenciones que describen cómo han de interpretarse estos documentos electrónicos, distinguiéndose entre la sintaxis y la semántica de dicho formato. Según establece Marcoux, la sintaxis sería el conjunto de reglas que se deben establecer en una secuencia de caracteres para que sea reconocida como un documento válido, mientras que la semántica sería el conjunto de

[13] Gary Cleveland. *Selecting Electronic Document Format* [documento HTML]. International Federation of Library Associations and Institutions, July 1999. Disponible en http://www.ifla.org/VI/5/op/udtop11/udtop11.htm (consultado el 3 de octubre de 2000).

reglas que permiten transformar un documento electrónico válido en un documento real o "verdadero"; pudiendo darse todo ello para formar un documento concreto (por ejemplo, un documento que será impreso en papel) o un documento abstracto, en el cual sólo existe un modelo conceptual para el mismo definido por la semántica del formato (por ejemplo, un modelo de documento hipertextual)[15].

La acción de marcar documentos tiene la función principal, como se ha comentado anteriormente, de describir el documento al cual se le está aplicando dicha acción. Esta descripción del documento puede realizarse a un nivel básico desde una doble perspectiva, lógica y física. La descripción lógica del documento determina los elementos lógicos o conceptuales que componen dicho documento (descripción interna). La descripción física hace mención al soporte material o elementos físicos que definen dicho documento (descripción externa)[16]. Así, tomando el ejemplo de un libro impreso, éste puede ser descrito lógicamente en una serie de tomos, donde cada tomo está integrado por una serie de capítulos y éstos, a su vez, por una serie de párrafos. Pero, de igual forma éste puede ser descrito físicamente, estableciendo para ello que este documento se compone de elementos tales como volúmenes, páginas, etc.

En el caso del documento electrónico se establece de igual modo esta doble descripción aunque con ligeras modificaciones debido principalmente a la menor relevancia del soporte material: el documento electrónico puede almacenarse en múltiples soportes magnéticos u ópticos sin que por ello varíen sus características básicas, pero necesita de un medio de presentación físico para poder ser visualizado por el ser humano (la pantalla del ordenador o el papel impreso, entre otros). En el entorno electrónico podemos, pues, describir

---

[14] Yves Marcoux. *Les formats normalisés de documents électroniques* [documento HTML]. Montreal: EBSI, Université de Montréal, 1999. Disponible en http://tornade.ere.umontreal.ca/~marcoux/grds/ico94.htm (consultado el 17 de noviembre de 2000).

[15] *Ibid.* http://tornade.ere.umontreal.ca/~marcoux/grds/ico94.htm

[16] François Role. "La norme SGML pour décrire la estructure logique des documents". *Documentaliste – Sciences de l'Information*, v. 28, nº 4-5, 1991, p. 187.

igualmente los elementos lógicos o estructurales que componen este documento. Pero la descripción física se ve algo alterada al ser ésta una forma de establecer en qué medio y de qué forma se va a presentar el documento. La **estructura** y el **formato** (o presentación) de los documentos electrónicos serán, por tanto, dos conceptos fundamentales que se manejarán asiduamente a lo largo de esta tesis doctoral.

Muchos de los sistemas informáticos para el tratamiento de documentos electrónicos, según palabras de André, Furuta y Quint, entienden y describen a éstos como una simple cadena de caracteres a los cuales se les añade información referente a cómo han de ser presentados ante un determinado medio de salida (pantalla del ordenador o impresora, principalmente, pero, por qué no, también medios auditivos); estos programas son conocidos por el nombre *procedural programming languages*[17]. Frente a este modo de operar, y según estos mismos autores, se encuentran los denominados *declarative programming languages*, orientados, en una primera fase, a la descripción de la estructura lógica del documento para, con posterioridad, establecer su correspondiente formato de presentación, o estructura física. Ambos conceptos de estructuras dentro de los documentos electrónicos serán analizados con mayor profundidad en posteriores secciones y capítulo de esta segunda parte.

La siguiente figura ilustra perfectamente esta dualidad en la descripción de un documento electrónico dado:

[17] J. André, R. Furuta, V. Quint (eds.). *Structured Documents*. New York [etc.]: Cambridge University Press, 1989, p. 3.

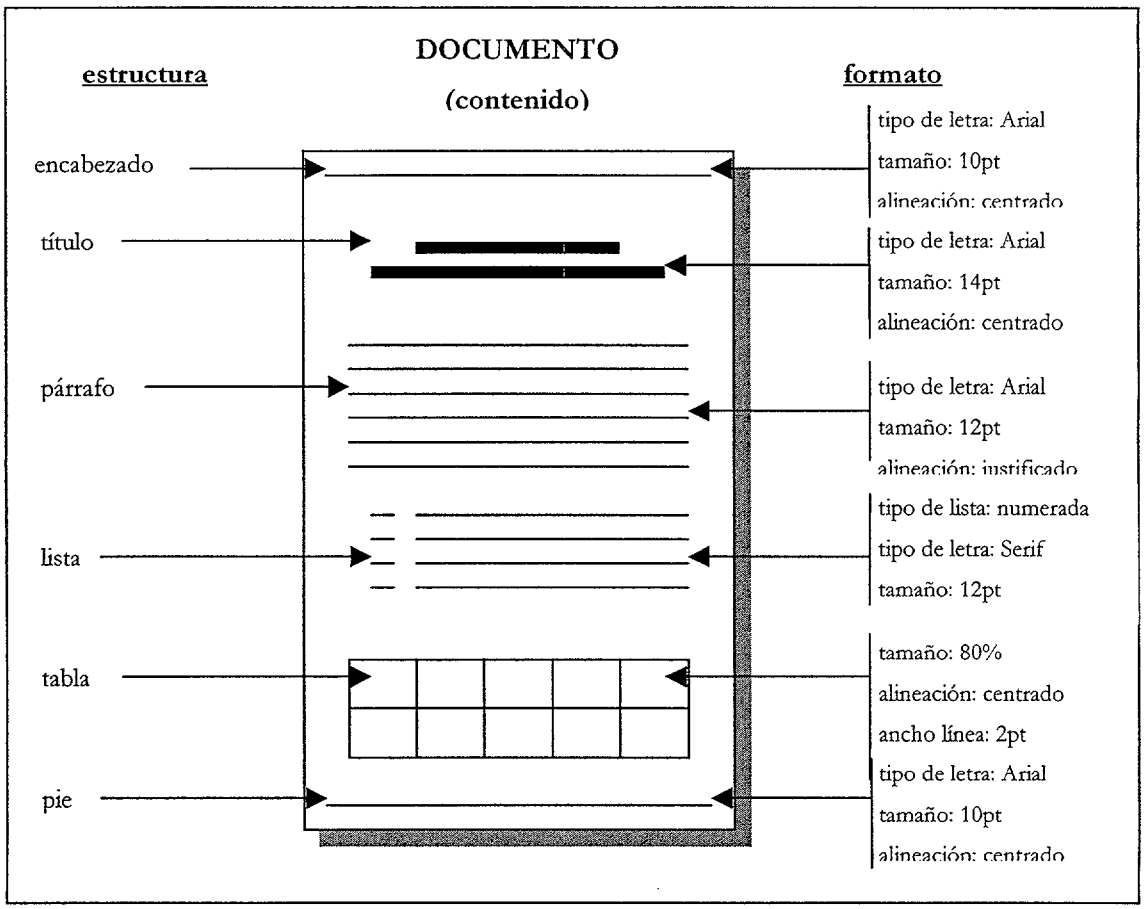


Figura II.1: definición estructura y definición de estilo en los documentos electrónicos.

El texto contenido en un documento electrónico, como lo señala Steven J. DeRose, puede ser entendido de múltiples maneras; a saber[18]: como un mapa de bits si el documento impreso es digitalizado, como una simple cadena de caracteres, como un conjunto formado por caracteres e instrucciones de procesamiento, como una representación fiel de la presentación final en un medio de salida, como una sucesión de objetos sin jerarquía alguna o, por último, como un modelo ordenado jerárquicamente de objetos contenedores. En este último caso, defendido en esta tesis como el mejor modelo para describir a los

documentos electrónicos, las partes esenciales de cualquier documento, a las cuales DeRose, denomina "objetos contenedores" (tales como párrafos, citas, frases enfatizadas, etc.) se integran y suceden en otros objetos contenedores mayores (como subsecciones, secciones, capítulos, etc.) siguiendo una lógica en dicha integración y sucesión de objetos: la denominada **estructura lógica** de los documentos. Todo ello conforma un modelo estructural jerarquizado de documento electrónico donde la estructura de cada documento podrá ser definida utilizando un tipo especial de lenguaje de marcado de texto, denominado marcado descriptivo, que será tratado ampliamente en apartados posteriores de este capítulo.

Con respecto a la utilización del marcado de texto por parte de los autores, L. Burnard define, de forma general, el proceso de marcado de documentos como la manera empleada para hacer explícita una interpretación del texto, siendo éste un proceso que dirige al usuario (hombre o máquina) hacia el modo en que debe ser interpretado el texto de un documento[19]. Pero sin duda, la definición de marcado más simple y ajustada es la que proporciona Ch. F. Goldfarb al indicar que el marcado es texto que ha sido añadido a los datos de un documento con la finalidad de proporcionar información sobre éstos[20]; o dicho de otro modo, el *marcado* es todo aquello que se encuentra en el texto de un documento que no es contenido propiamente dicho pero que ayuda a definir ciertas características del mismo. La plasmación física de este proceso de marcado de textos en el entorno electrónico se realiza mediante la inclusión de unas instrucciones determinadas conocidas por el nombre de códigos o, más frecuentemente, etiquetas (*tags* o *labels*, en su acepción anglosajona). Estos códigos o etiquetas vienen normalmente representados por caracteres

---

[18] Steven J. DeRose, David G. Durand, Elli Mylonas, Allen H. Renear. "What is Text, Really?". *Journal of Computing in Higher Education*, v. 1, nº 2, 1990, p. 8 y posteriores.

[19] Lou Burnard. "Markup and Markup Languages" [documento HTML]. En: *What is SGML and How Does it Help?* Oxford: The Humanities Computing Unit, Oxford University Computing Service, 1999. Disponible en http://www.hcu.ox.ac.uk/TEI/Papers/EDW25/W25C.htm (consultado el 5 de junio de 2000).

[20] Charles F. Goldfarb, Yuri Rubinsky. *The SGML Handbook*. [1ª ed., 4ª reimp.] New York [etc.]: Oxford University Press, 1990, p. 21.

propios de los teclados convencionales de los ordenadores, por lo que su inserción en el documento se puede realizar con simples editores de texto.

Del marcado de los documentos electrónicos se inferirá algo de vital importancia, tal y como se explicará posteriormente: la identificación de elementos estructurales dentro de los mismos. En principio, tal y como señalan Y. Marcoux y M. Sévigny, todo documento electrónico, entendido como un fichero informático, no contiene una estructura inherente más que la secuencia lineal de una cadena de caracteres (o *bytes*). Si algunas partes del documento pueden ser identificadas, es posible establecer un mecanismo o convención para denotar dichas partes. Por ejemplo, se pueden establecer localizaciones fijas para dichas partes junto al documento electrónico, o también implementar un sistema de punteros y/o separadores de dichas partes. Pero el método más eficaz resulta el marcado (*markup*) o etiquetado (*tagging*) de las mismas. El marcado consiste en insertar dentro del documento electrónico pequeñas cadenas de caracteres, o etiquetas, para identificar el comienzo y el final de una determinada parte del documento. El conjunto de etiquetas que se encuentran dentro de un documento electrónico es conocido también habitualmente como marcado (*markup*)[21].

Las reglas que establecen qué marcas han de ser empleadas, de qué modo se distinguirá la marca del texto del documento y cómo se insertarán éstas (la gramática y su sintaxis), y cuáles son las marcas permitidas en cada una de las partes del texto, es lo que se conoce como un **lenguaje de marcado** (*markup language*)[22]. No es, por tanto, más que un conjunto

[21] Yves Marcoux, Martin Sévigny. "Why SGML? Why Now?". *Journal of the American Society for Information Science*, v. 48, nº 7, 1997, p. 586.

[22] Algunos autores castellanoparlantes traducen también este término anglosajón como "lenguaje de etiquetado" o "lenguaje de marcas". Cualquiera de estas acepciones la consideramos como válida. Pero, para los efectos de esta tesis doctoral usaremos el término de "lenguaje de marcado" debido a que, por un lado, tradicionalmente se ha venido utilizando esta traducción del término inglés y, de otra parte, hace mención al proceso en sí de asignación de marcas, tal y como se explicó al principio de este capítulo. En cualquier caso, la traducción del término inglés "markup language" no siempre es fácil, y a veces sesgada o incompleta, dependiendo en gran medida del contexto en el que nos movamos. Tal es el caso, por ejemplo, del famoso diccionario multilingüe de O. Voonhals, que traduce dicho término a nuestro idioma como "lenguaje de señalización tipográfica". Otto Vollnhals. *Multilingual Dictionary of Electronic Publishing*. München [etc.]: K. G. Saur, 1996, p. 105.

de marcas o etiquetas normalizadas, y más o menos flexibles, utilizadas para codificar textos siguiendo unas determinadas reglas gramaticales y sintácticas.

En resumen, un documento marcado es aquel que incluye dos tipos de información: el texto, el contenido o los datos propiamente dichos y las etiquetas que se han empleado para marcarlo con un determinado fin.

J. H. Coombs y otros colaboradores establecieron en los años 80 una clasificación de los tipos o técnicas de marcado de texto, que ha sido seguida y citada por numerosos especialistas en la materia. Algunas de las técnicas más características serán analizadas con mayor profundidad en apartados posteriores. Estos autores contemplan seis tipos diferentes de técnicas de marcado de texto[23]:

1. **Marcado de puntuación** (*punctuational*): consiste en el uso de un conjunto cerrado de marcas o signos para proporcionar, principalmente, información sintáctica sobre las expresiones escritas. Este tipo de marcado ha sido utilizado por la humanidad a lo largo de los siglos y es considerado parte sustancial del proceso de escritura. Desde este punto de vista, sólo algunos manuscritos antiguos no llevan ningún tipo de marcado (aquellos que se han escrito sin utilizar signos de puntuación, como eran los *scripta continua*).

2. **Marcado de presentación** (*presentational*): un mayor nivel de codificación del texto permite a los autores establecer presentaciones más ricas y bellas en diferentes medios de publicación. Para ello se establecen marcas que indican, por ejemplo, el espacio horizontal entre palabras y el vertical entre las líneas, los saltos de párrafo, la numeración para las listas, la paginación, etc. El marcado orientado a la presentación

[23] James H. Coombs, Allen H. Renear, Steven J. DeRose. *Op. cit.*, http://www.oasis-open.org/cover/coombs.html

clarifica, por tanto, la presentación del documento, haciéndolo más conveniente para su lectura por parte del ser humano.

3. **Marcado de procedimiento** (*procedural*): con la integración y uso de los procesadores automáticos de textos electrónicos el marcado orientado a la presentación es sustituido, aunque en estrecha e íntima relación, por el marcado de procedimiento. Este marcado consiste en la aplicación, de forma directa o indirecta por parte del usuario, de una serie de comandos informáticos para indicar a la aplicación informática cómo debe ser formateado o codificado el texto electrónico. En otras palabras, las instrucciones del marcado de procedimiento indican a una determinada aplicación informática que haga algo con un determinado texto (por ejemplo, establecer diferentes espaciados de línea para diferentes párrafos, cambiar la sangría de una determinada lista, establecer un control de viudas y huérfanas en los párrafos, etc.). Cada aplicación informática de procesamiento electrónico de texto (los comúnmente llamados procesadores de texto) aplica un determinado lenguaje de marcado propietario para la codificación, formateado y presentación de los textos electrónicos.

4. **Marcado descriptivo** (*descriptive*): este marcado permite a los autores identificar los tipos o piezas estructurales que componen el texto. La principal diferencia entre el marcado de procedimiento y el descriptivo radica en que el establecimiento de marcas por este último hace mención a la naturaleza de los elementos del texto, sin tener en cuenta cómo han de ser formateados éstos para su presentación. En otras palabras, el marcado descriptivo declara que un determinado segmento del documento pertenece o es miembro de una determinada clase particular. Para la creación de documentos electrónicos con un marcado descriptivo se pueden emplear herramientas informáticas para el procesamiento del texto que, en un principio, van destinadas exclusivamente al marcado de procedimiento. Este hecho permite que el marcado descriptivo pueda ser procesado por un número de aplicaciones de distinta índole.

5. **Marcado referencial** (*referential*): El marcado referencial, como su propio nombre indica, es el que se emplea para hacer referencia a entidades externas al documento, el cual será reemplazado por éstas durante el procesamiento del documento en cuestión. Normalmente estas entidades suelen hacer referencia a imágenes electrónicas, gráficos, sonidos, otros documentos electrónicos, etc., esto es, aquello que no es codificado como texto del documento propiamente dicho. Estas entidades suelen estar almacenadas en ficheros electrónicos distintos al del documento, e incluso en diferentes ordenadores. La mayoría de los actuales procesadores de texto tienen la capacidad funcional de incorporar marcado referencial a través de la definición de variables por parte del usuario, la inserción directa en el documento de estos ficheros externos o a través de una serie de comandos.

6. **Metamarcado** (*metamarkup*): un lenguaje de metamarcado, o también denominado más comúnmente como metalenguaje de marcado de texto, permite a los autores controlar la interpretación del marcado del documento y extender o ampliar el vocabulario de los lenguajes de marcado descriptivo. El metamarcado suele aparecer en la forma de *declaraciones de marcado*, las cuales establecen la forma, uso y reglas, para un conjunto de marcas descriptivas establecidas de antemano por el usuario. En definitiva, se trata de un lenguaje que permite crear otros lenguajes de marcado descriptivo, siendo SGML, como se verá posteriormente, el mejor ejemplo dentro de esta categoría.

En los apartados siguientes se mostrará con mayor profundidad la evolución y aplicación de algunos de estos lenguajes de marcado de texto electrónico, en especial aquellos que han tenido una mayor trascendencia. Realizando una primera aproximación al tema, en estos lenguajes de marcado de texto electrónico vamos a encontrar una progresiva transición en el uso de la marca, que irá desde su uso exclusivo para la descripción de formatos de presentación hasta la descripción de elementos estructurales contenidos en el documento.

II.1.2.2. LENGUAJE DE MARCADO DE FORMATO, PRESENTACIÓN O PROCEDIMIENTO:

En los albores de las ciencias de la computación, en los años 40 y 50, los ordenadores eran empleados principalmente para procesar grandes cantidades de datos simples, numéricos o alfanuméricos, y realizar cálculos de forma automática, restringiéndose su utilización a laboratorios de tecnología de grandes instituciones y corporaciones públicas y privadas. Con la automatización en la producción de los textos por parte de los ordenadores con capacidades para soportar software de procesamiento de texto y, por tanto, capaces de procesar y automatizar partes del proceso de creación y edición de un documento, el concepto de marca se hizo extensible a todos los tipos especiales de códigos de marcado que se insertaban dentro de los documentos electrónicos a través de dichos programas informáticos. La tipografía informática permitía a los autores redactar el texto del documento y establecer el tipo de formato que se deseaba para el mismo. De este modo, el ordenador establecía que la reproducción de dicho documento estaba sujeta a un determinado formato de archivo electrónico, compuesto por la combinación del texto del documento y los códigos introducidos que especifican el formateado del documento en sí y partes específicas del mismo (lo que debe ir negrita, en cursiva, subrayado, etc.). Por último, el sistema informático convierte este formato de reproducción de documentos electrónicos en signos capaces de ser percibidos e interpretados por el ser humano ante un determinado medio de presentación, bien la pantalla electrónica del ordenador (con una determinada interfaz de usuario) o bien el papel en dispositivos de salida de impresión. Este modelo de trabajo con documentos electrónicos tendrá su máxima expresión con los programas informáticos de procesamiento de texto WYSIWYG (*What You See Is What You Get*). Aunque los primeros procesadores de texto de este tipo para ordenadores personales aparecen en los años 70[24], el primer programa de autoedición de cierta entidad fue lanzado

[24] El primer ordenador personal del mundo que incorporaba un editor WYSIWYG dentro de las aplicaciones residentes fue "The Alto" de la compañía Xerox Corporation, lanzado a principio de dicha década. Para una

en 1985 y llevaba el nombre de *Aldus PageMaker*[25]. Estos programas presentan una interfaz de usuario completamente "amigable" para el creador de documentos electrónicos ya que plasman en la pantalla del ordenador la presentación de dicho documento tal y como será su acabado impreso en papel. Ejemplos representativos de este tipo de procesadores de texto son bien conocidos por todos los usuarios, como son el caso de *WordPerfect*, *Lotus AmiPro*, *Adobe PageMaker* o *Microsoft Word*. Este modelo de tratamiento del texto de los documentos electrónicos en el cual el formato de reproducción se limita exclusivamente a describir la presentación de dichos documentos utiliza un tipo de notación tipográfica denominado *marcado de formato*.

Un **lenguaje de marcado de formato**, también denominado lenguaje de marcado de procedimiento (*procedural markup language*), lenguaje de marcado basado en la apariencia o presentación (*Appareance-based Markup Language*) o lenguaje de descripción de página (*Page Description Language*), es aquel que especifica cómo debe ser procesado el contenido del documento electrónico para asignarle una determinada presentación entendible por el usuario. Así, se insertarán marcas o códigos en el documento electrónico (codificación) de una forma implícita (por el programa informático, a través de una serie de comandos u órdenes establecidas por el usuario) o explícita (directamente por el usuario) para indicar a la aplicación informática (normalmente un procesador de texto) cómo debe ser procesada dicha codificación del texto del documento electrónico y, por consiguiente, establecer una determinada presentación del mismo ante un determinado medio de salida (pantalla o papel impreso, normalmente). Con este tipo de marcado se especifican aspectos tales como la disposición del texto en las páginas del documento, la fuente o fuentes de caracteres a emplear, palabras en negrita, cursiva y subrayadas, y otras muchas características

mayor información véase el documento sobre la historia y desarrollo de los productos de esta compañía informática en http://www.parc.xerox.com/hist-lst.html (consulta el 25 de septiembre de 1999).

[25] Para una mayor profundidad sobre el tema de los procesadores de texto WYSIWYG recomendamos el artículo de Conrad Taylor. "What has WYSIWYG done to us?" [documento HTML]. *The Seybold Report on Publishing Systems*, v. 26, nº 2, 30 de septiembre de 1996. Disponible en http://www.ideography.co.uk/library/seybold/WYSIWYG.html (consultado el 10 de agosto dl 2000).

meramente tipográficas[26]. Las marcas empleadas para la codificación del texto del documento electrónico toman el nombre de **códigos de formateado** (*formatting codes*), los cuales van mezclados con el texto original del documento[27].

Desde su origen, las marcas o códigos que conforman un determinado lenguaje de este tipo han venido siendo específicos de un determinado sistema informático de procesamiento de texto electrónico. Así, muchas compañías comerciales de software editor de textos lanzaba al mercado un determinado programa con un lenguaje de marcas con significado exclusivo para dicho sistema, y en muchos casos entendible solamente por un determinado sistema operativo y procesable en un *hardware* concreto. Aunque en los últimos tiempos estos programas de procesamiento de texto electrónico se han hecho más "inteligentes", llegando a entender diversos lenguajes de marcado de formato propietarios de diversas compañías informáticas, no dejan de ser eso, lenguajes propietarios, sujetos a las variaciones que la compañía que ostenta sus derechos legales y comerciales desee establecer. Por otro lado, y en un principio, estos códigos de formateado eran insertados por el propio usuario a través del teclado del ordenador y visibles en todo momento; se podía distinguir perfectamente lo que era código del contenido textual propiamente dicho del documento. Los modernos programas de procesamiento de textos, con interfaces WYSIWYG, ocultan estas marcas al usuario, mostrando en pantalla únicamente el resultado final de dicho procesamiento; esto es, el modo en que será presentado finalmente el documento en el medio de publicación elegido (normalmente, en papel impreso de tamaño DIN-A4). Las siguientes figuras ilustran la codificación del texto a través del lenguaje de marcado RTF (*Rich Text Format*), el cual será explicado posteriormente, y cómo es presentado dicho documento en un procesador de texto del tipo WYSIWYG.

---

[26] Marcello P. Bax. *Introdução às Linguagens de Marcas* [documento HTML]. PARADIGMA Internet, 14 de abril de 2000. Disponible en http://www.paradigma.com.br/XML/introxml.htm (consultado el 17 de agosto de 2000).

[27] *SGML: Getting Started* [documento HTML]. Arbortext, 1995. Disponible en http://www.arbortext.com/Think_Tank/SGML_Resources/Getting_Started_with_SGML/getting_started_with_sgml.html (consultado el 17 de agosto de 2000).

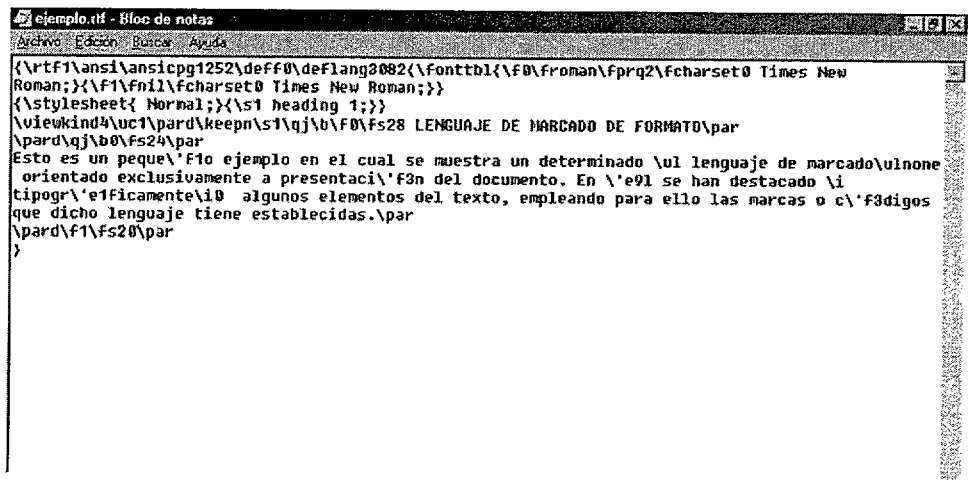


Figura II.2: Texto de un documento electrónico codificado en formato RTF.

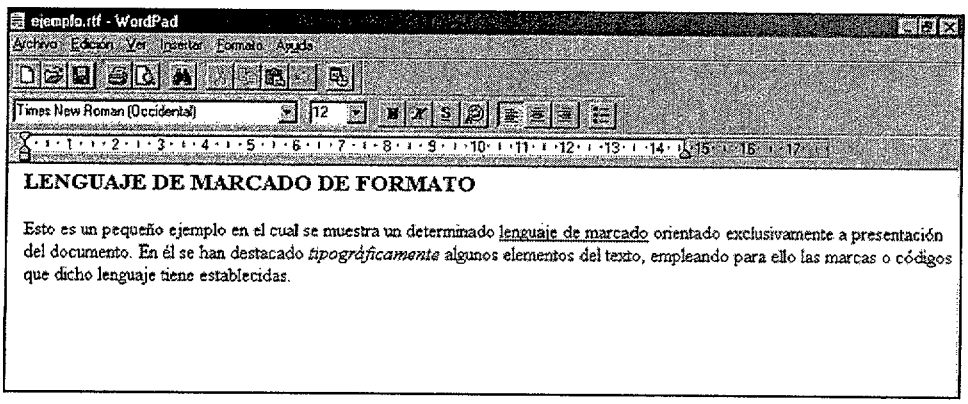


Figura II.3: El mismo documento tratado con un procesador WYSIWYG.

Sin entrar en demasiadas consideraciones al respecto, podemos señalar algunos de los formatos de lenguaje de marcado propietarios más conocidos y ampliamente utilizados por los usuarios para el tratamiento de textos electrónicos[28]. Algunos de éstos son los siguientes:

[28] A pesar de no contemplar todos los formatos para la edición de textos electrónicos expuestos a continuación, resulta especialmente interesante por su valor descriptivo e histórico la lectura del artículo elaborado por P. Hípola y R. Eito Brun, al cual remitimos al lector para una mayor profundidad en el tema

- **Nroff** (*Next run-off*) / **Troff** (*Typesetter run-off*) : Basándose en el comando *roff* del sistema operativo UNIX para la edición y procesamiento de textos electrónicos, Joe Ossanna, a principio de los años 70, aumentó las capacidades de este comando, pasando a denominarlo nroff. A medida que fue aumentando en potencialidad en los siguientes años pasó a denominarse *troff*, dando soporte a los sistemas gráficos informatizados de tipografía, incorporando múltiples tipos de fuentes, tamaños y caracteres adicionales que las impresoras habituales de aquellos tiempos no soportaban. El comando *troff* es utilizado para procesar un documento que ha de ser visualizado en un terminal gráfico, en una fotocomponendora o en una impresora láser, mientras que el comando *nroff*, su predecesor, era utilizado para producir un documento de salida en una impresora matricial o en una pantalla de mapa de caracteres. La entrada de documentos para estas aplicaciones UNIX se basa en la introducción de texto electrónico con comandos de formateado que indican el estilo de las fuentes de texto y la organización de dicho texto en un dispositivo de salida[29]. Este formato aún se sigue empleando en algunas organizaciones.

- **PostScript:** el lenguaje de descripción de páginas *PostScript* fue desarrollado por John Warwock y Chuck Geschke, fundadores de la compañía Adobe Systems Incorporated. Este importante formato fue introducido por dicha compañía en 1985 en la impresora Apple LaserWriter. El propósito principal de PostScript era el de proporcionar un lenguaje adecuado para describir imágenes en un dispositivo de manera independiente, describiendo para ello las imágenes sin referencia alguna a rasgos o características del dispositivo (por ejemplo, la resolución de la impresión). Aunque se trata, en un principio, de un formato propietario, debido a su enorme éxito y aceptación fue convirtiéndose en un estándar de facto en el entorno de la impresión digital. Se trata,

del empleo y repercusión de estos lenguajes orientados al formato frente a los posteriores lenguajes descriptivos, analizados con posterioridad en este capítulo de la tesis doctoral. Véase, Pedro Hípola, Ricardo Eito Brun. "Edición digital: formatos y alternativas". *El Profesional de la Información*, v. 9, nº 10, octubre 2000, pp. 4-14.

pues, de un potente lenguaje que traslada textos y gráficos contenidos en los documentos electrónicos a dispositivos de salida de impresión de alta calidad, como son las impresoras láser[30].

- **PDF** (*Portable Document Format*): anunciado en 1992 por la compañía Adobe, puede ser definido como un "PostScript con hipertexto"[31]; esto es, un formato PostScript modificado con potentes capacidades hipertextuales para ser tratado por el *software* Adobe Acrobat (tanto editor WYSIWYG como visualizador de documentos de este tipo). Como se señala en el manual oficial de este desarrollo de la compañía Adobe, PDF es un formato de fichero electrónico utilizado para representar un documento de manera independiente de la aplicación de *software*, *hardware* y sistema operativo que lo creó[32]. Un fichero PDF contiene un documento PDF y otros datos de soporte. Se trata de uno de los lenguajes de marcado de documentos electrónicos más conocidos y utilizados actualmente en el mundo de la edición electrónica de documentos. Aunque se trata, como decíamos, de un formato orientado principalmente al formateado de documentos electrónicos, este lenguaje incluye ciertas capacidades que le permiten definir la estructura de los grandes bloques temáticos en los que se descompone el

[29] Jan Pardoe. *UNIX Text Formatting Using the -ms Macros* [documento HTML]. Berkeley, California: Computing Services Library, University of California, 15 de marzo de 1995. Disponible en http://www.cs.berkeley.edu/~janp/Help/textms.html (consultado el 25 de septiembre de 2000).

[30] Peter Weingarter. *A First Guide to PostScript* [documento HTML] 18 de enero de 1997. Disponible en http://www.cs.indiana.edu/docproject/programming/postscript/postscript.html (consultado el 26 de agosto de 2000).

[31] Judith Wusteman. "Formats for the Electronic Library" [documento HTML]. *Ariadne: the Web version*, nº 8, march 1997. Disponible en http://www.ariadne.ac.uk/issue8/electronic-formats/intro.html (consultado el 12 de noviembre de 1999).

[32] *Portable Document Format. Reference Manual, version 1.3* [documento PDF]. Adobe Systems Incorporated, 11 de marzo de 1999. Disponible en http://partners.adobe.com/asn/developer/acrosdk/DOCS/pdfspec.pdf (consultado el 26 de agosto de 2000).

documento, hecho éste que le hace estar, en realidad, a caballo entre los lenguajes orientados a la presentación y los lenguajes de marcado descriptivo[33].

- **RTF** (*Rich Text Format*): desarrollado por la empresa Microsoft Corporation como un lenguaje de formateado de textos y gráficos electrónicos de aplicación y validez en todas las aplicaciones informáticas de esta compañía (Word, Excel, Access, etc.) para el intercambio de información basada en el formato[34]. La pujanza en el mercado de los productos de esta compañía ha sido tal que en la actualidad este lenguaje propietario es anunciado por la propia compañía Microsoft, como un formato para el intercambio de texto y gráficos electrónicos que puede ser utilizado con diferentes dispositivos de salida y en otros sistemas operativos, como en los de las compañías IBM o Apple Macintosh. RTF utiliza el juego de caracteres ANSI, PC-8, Macintosh o de los ordenadores IBM para controlar la representación y el formateado del documento electrónico, tanto en pantalla como en impresora.

[33] No es éste el único caso que se encuentra en esa frontera a veces difusa entre lenguajes de marcado orientados al formato y los lenguajes de marcado estructural. Por ejemplo, un lenguaje que no tuvo prácticamente ninguna repercusión en el ámbito comercial pero sí en el entorno de la investigación científica por sus influencias tanto en los actuales programas de edición para ordenadores de sobremesa así como en los diversos lenguajes de marcado descriptivo que han venido apareciendo, fue el denominado SCRIBE. Este interesante pero poco conocido lenguaje de marcas fue desarrollado por Brian Reid de la Universidad de Carnegie Mellon en Pittsburg a finales de la década de los 70 y presentado en 1980 como parte integrante del desarrollo práctico de su tesis doctoral en (Brian K. Reid. *Scribe: A Document Specification Language and its Compiler.* Ph. D. Dissertation, Carnegie Mellon University, Pittsburg, PA. October, 1980) Este lenguaje de descripción de página permitía crear documentos electrónicos estructurados con la posibilidad de ser formateados de forma independiente por parte de las primeras impresoras láser aparecidas en aquellos años. Para una mayor información, véase el artículo de Karen A. Frenkel. "Profiles in computing: Brian K. Reid: a graphic tale of a hacker tracker" [documento PDF]. *Communications of the ACM*, v. 30, n° 10, 1987. Disponible en http://www.acm.org/pubs/articles/journals/cacm/1987-30-10/p820-frenkel/p820-frenkel.pdf (consultado el 29 de enero de 2001).

[34] *Rich Text Format (RTF) Specification and Sample RTF Reader Program, Version 1.5* [documento HTML]. Disponible en http://msdn.microsoft.com/library/specs/richtextformatrtfspecificationsamplertfreaderprogramversion15.htm (consultado el 21 de septiembre de 2000).

Este tipo de lenguajes, a pesar de los beneficios reportados al mundo de la edición electrónica de documentos, cuenta con una serie de desventajas desde la óptica de la gestión de la información en entornos informatizados, a saber[35]:

- Son sistemas propietarios de producción y edición de textos electrónicos, lo que implica que para poder manejar uno de estos lenguajes se debe contar con el *software* apropiado que la compañía propietaria de los derechos haya desarrollado. Por otro lado, muchos de estos lenguajes van normalmente ligados a un determinado sistema operativo y a un determinado *hardware*. Todo ello imposibilita en gran medida la transferencia de información dentro de las organizaciones debido a las grandes dificultades para desarrollar sistemas que compatibilicen diferentes formatos electrónicos de representación de documentos.

- No aportan ninguna información de tipo semántico sobre el contenido del documento por lo que no pueden ser utilizados en aplicaciones de recuperación de información electrónica. Los módulos de interrogación y búsqueda documental de dichas aplicaciones informáticas se ven, por tanto, incapaces de discernir cuáles son las partes del documento en las que se encuentra el texto que contiene la mayor carga semántica o descriptiva del contenido informativo de dicho documento.

- No registran la estructura lógica del documento electrónico. Tan sólo, a través de un proceso de abstracción realizado por el usuario al visualizar en pantalla o en papel la apariencia del documento en cuestión, puede deducir dicha estructura.

- Es un lenguaje muy poco flexible debido a que cualquier cambio que desee realizar el usuario sobre el estilo del documento implicará invariablemente tratar nuevamente el

[35] Ch. F. Goldfarb, Y. Rubinsky. *Op. cit.*, p. 7.

proceso de marcado del documento en cuestión para reflejar dichos cambios. Cada nueva modificación en el estilo conlleva la apertura y reformateado del documento.

- Es un proceso que consume una gran cantidad de tiempo y esfuerzo, en especial en aplicaciones profesionales de edición de alta cualificación, siendo, por tanto, frecuente la inclusión de errores de marcado a la hora de asignar el formato adecuado.

Frente a esta idea de utilización del marcado de documentos electrónicos para asignar un determinado formato de presentación al documento y a ciertas partes del texto contenido en él, surge en los años 60, y de forma paralela al desarrollo y evolución de los lenguajes de marcado orientados al formato, una corriente de opinión crítica a la conveniencia de estos sistemas para un eficaz tratamiento de los documentos electrónicos dentro de las organizaciones: frente a la idea de un marcado de los documentos orientado a la presentación visual (estilistas) surgirá la idea de la primacía del marcado para la definición de estructuras lógicas en estos documentos (estructuralistas).

II.1.2.3. CODIFICACIÓN GENÉRICA:

A medida que la producción de textos electrónicos en las organizaciones aumentaba de forma exponencial, los redactores de estos textos empezaron a exigir mayores funcionalidades a los sistemas informatizados para el procesamiento de los mismos. La asignación de marcas de formato a los documentos electrónicos se había vuelto en muchos casos una labor ardua, lenta y tediosa, por lo que se hacía necesario desarrollar algún sistema que simplificase y agilizase dicha tarea. Los lenguajes de programación permitían crear pequeños programas, conocidos por el nombre de *macros*, que automatizaban el proceso de asignación de marcas al documento. Las macros eran llamadas (*macro calls* o

format calls) para identificar las partes del documento donde el procesador de texto debía actuar, estableciendo las correspondientes marcas de formateado. Dado que estas partes del documento (título, encabezamientos, párrafos, etc.) eran comunes en su formato o estilo de presentación con respecto a otros documentos similares, parecía lógico suponer que estas macros deberían llevar nombres fácilmente identificables por los usuarios. Así, en vez de denominar a una macro como un procesamiento específico del tipo "format-17", sería más lógico asignarle el nombre del elemento del documento al cual fuese a afectar. En este caso tomaría el nombre de "heading", dado que aplicaría un determinado formateado a los encabezamientos de los documentos electrónicos a los cuales se fuese a aplicar esta macro. De este modo, y ésta fue la gran aportación, los usuarios tendían a otorgar nombres significativos a las etiquetas, reconociendo así el predominio de elementos estructurales frente al formateado del texto. A cada una de estas denominaciones nemotécnicas de las macros se le dio el nombre de **identificador genérico** (*generic identifier*), y al sistema global de marcado de documentos electrónicos a través de llamadas de macros con identificadores genéricos se le denominó **codificación genérica** (*generic coding*) o etiquetado generalizado (*generalized tagging*)[36]. Muchos autores señalan a William Tunnicliffe, presidente a finales de los años 60 de la *Graphic Communications Association* (GCA), como el precursor de la codificación genérica al dar a conocer en septiembre de 1967 dichos postulados en la *Canadian Government Printing Office*[37].

La codificación genérica dentro de los lenguajes de procesamiento de texto electrónico orientados al formato o presentación supone un paso más allá en el proceso evolutivo de los lenguajes de marcado. Un paso más allá dado, aunque su fin principal sigue girando en torno al formato de presentación de los documentos electrónicos, en el cual ya se empieza a diferenciar ciertas partes estructurales contenidas en los documentos. Así es, la utilización

[36] Ch. F. Goldfarb, Y. Rubinsky. *Op. cit.*, p. 239.

[37] SGML Users' Group. *A Brief History of the Development of SGML* [documento HTML]. 11 de junio de 1990. Disponible en http://www.oasis-open.org/cover/sgmlhist0.html (consultado el 25 de agosto de 2000).

de los identificadores genéricos supone una primera aproximación a la diferenciación entre contenido y formato dentro de los documentos electrónicos.

B. Marchal destaca como beneficios principales aportados por este modelo de codificación genérica, frente al lenguaje de marcado orientado al formato, los dos siguientes[38]:

- Se consigue una mayor "portabilidad" y se aumenta la flexibilidad del documento: ahora no es necesario editar el documento electrónico para establecer alguna modificación en la presentación del mismo dado que para cambiar esta apariencia basta con editar la macro y adaptarla a nuestras exigencias. El cambio producido en la macro se transfiere de forma automática al formato de presentación del documento.
- El marcado del documento se reorienta hacia la descripción de la estructura interna del mismo, al ser un inicio de identificación de piezas lógicas básicas.

De forma parecida se expresa A. Brüggemann-Klein, señalando que la separación de las especificaciones de diseño del texto del documento reporta tres importantes ventajas para la edición y publicación de documentos electrónicos:[39]

- El documento electrónico puede ser formateado según diversas especificaciones de diseño sin que por ello se vea modificado el contenido original del documento.
- Aquellos documentos que tienen una misma organización estructural pueden compartir una especificación de diseño común que imprima consistencia o coherencia a la presentación de dichos documentos.

---

[38] Benoît Marchal. *XML by Example.* Indianapolis, Indiana: Que, 2000, p. 17.

[39] Anne Brüggemann-Klein. *Formal Models in Document Processing* [documento PostScript]. Freibrug: Informatik and der MathematischenFakultät, Albert-Ludwigs-Universität, Freiburg 1993, p. 7. Disponible en ftp://ftp.informatik.uni-freiburg.de/documents/papers/brueggem/habil.ps (consultado el 6 de noviembre de 2000).

- Al separar las cuestiones relacionadas con el diseño y presentación del documento se le está otorgando, asimismo, una mayor importancia al desarrollo y claridad de dichas tareas.

El primer proyecto de envergadura basado en la codificación genérica se desarrolló a mediados de la década de los 60 y fue conocido por el nombre de **GenCode**. Stanley Rice, prestigioso diseñador y maquetador norteamericano de libros, expuso la idea de construir un formato electrónico universal para el intercambio de textos entre empresas editoriales de diversas partes del mundo, basándose para ello en un marcado de tipo genérico. El director de la *Graphic Communications Association* (GCA), Norman Scharpf, se hizo eco de esta idea, poniendo en marcha un ambicioso proyecto con la finalidad de crear un modelo normalizado de codificación genérica de aplicación en el campo de la composición tipográfica en un medio electrónico. La idea básica subyacente era la de desarrollar un conjunto de códigos genéricos que permitiera a las empresas del sector enviar datos a sus clientes evitando gastos de conversión de formatos electrónicos y pérdidas de tiempo cada vez que se cambiase de sistemas y plataformas informáticas. Este modelo, por tanto, debería construirse basándose en el procesamiento de los documentos electrónicos en un formato abierto de publicación, no propietario, que pudiera ser fácilmente intercambiable y manipulable por cualquier sistema informático existente en esos momentos en el mercado[40]. El comité que se encargó de está tarea desarrolló el llamado *GenCode concept*, que establecía, por un lado, la necesidad de diferentes códigos genéricos para distintos tipos de documentos, y añadía, además, la idea de que los documentos de menor tamaño podían ser incorporados como elementos dentro de los documentos más extensos[41].

[40] Liora Alschuler. *ABCD... SGML: A User's Guide to Structured Information.* London [etc.] : International Thomson Computer Press, 1995, p. 6.

[41] Aunque en estos momentos no se puede hablar todavía de definición de tipos documentales (DTDs), tal y como serán entendidos en SGML y, posteriormente, en XML, no es menos cierto que ya se empezaba a tener en cuenta este elemento tan característico.

Otro importante desarrollo en este campo lo constituyeron los lenguajes de marcado TeX y LaTeX.

TeX fue otro sistema de tipografía informática basado en macros desarrollado por Donald E. Knuth a finales de los años 60, aunque en principio orientado exclusivamente al marcado procedimental de documentos electrónicos. Kunt, profesor de matemáticas y ciencias de la computación, decidió crear material docente para su asignatura en forma de libro electrónico viendo, para ello, la necesidad de implementar un lenguaje de marcado nuevo que satisficiera la necesidad de incluir, y, por tanto, formatear, ecuaciones matemáticas. TeX es un procesador de macros, independiente de la máquina y el sistema operativo que lo procese, que ofrece unas potentes capacidades de programación, pero que a la vez esto mismo le hace ser un lenguaje complejo y difícil de manejar. En su implementación básica, TeX contemplaba más de 300 controles y macros. El propio Knuth desarrolló un paquete de macros, conocido por el nombre de *plainTEX*, que podían ser empleadas con este lenguaje y que simplificaban notablemente la aplicación de este sistema[42].

LaTeX es un paquete informático de macros, escrito originalmente por Leslie Lamport, que mejora y simplifica la generación de textos electrónicos de su predecesor al proporcionar un sistema de procesamiento de documentos. LaTeX se concentra, en un principio, en la descripción básica de la estructura lógica del documento (título, encabezamientos, párrafos, etc.), sin tener en cuenta la presentación del mismo, pues de ello se ocuparía otro lenguaje más apropiado para tal fin. Así, el sistema LaTeX requería que el marcado del texto que definía la estructura lógica básica se realizase en un fichero ASCII con códigos LaTeX para, posteriormente, convertir el fichero codificado con este lenguaje a un formato de presentación apropiado, normalmente en PostScript[43]. Durante varias décadas LaTeX se convirtió en el lenguaje de creación de documentos electrónicos preferido y más utilizado por la comunidad de científicos e ingenieros, especialmente por

---

[42] *TeX Frequently Asked Questions, version 2.4.7a* [documento HTML]. UK TeX Users Group, 7 de Julio de 2000. Disponible en http://www.tex.ac.uk/cgi-bin/texfaq2html (consultado el 24 de agosto de 2000).

[43] J. Wusteman. *Op. cit.*, http://www.ariadne.ac.uk/issue8/electronic-formats/intro.html

matemáticos e informáticos, al proporcionar unas magníficas capacidades para la representación de fórmulas matemáticas en los documentos.

La codificación genérica constituyó una etapa crucial en la construcción de sistemas automatizados de tratamiento de textos electrónicos, como se verá posteriormente, al reflejar la estrecha relación que existe entre los atributos del documento y su procesamiento. A pesar de esta considerable mejora con respecto a los lenguajes orientados exclusivamente a la representación del texto, la codificación genérica era conceptualmente insuficiente debido a la imposibilidad de definir satisfactoriamente ciertos aspectos de los documentos electrónicos complejos, como la asignación de identificadores específicos a algunos elementos del documento, la representación de la estructura jerárquica que conforman los elementos de un documento, etc.

El modelo de lenguaje de marcado de texto electrónico surgido para hacer frente a estos problemas, el lenguaje de marcado descriptivo, tomará y ampliará algunas de las ideas propuestas por la codificación genérica y la utilización de identificadores genéricos para marcar ciertas partes estructurales existentes en los documentos electrónicos, tal y como se verá a continuación.

II.1.2.4. LENGUAJE DE MARCADO DESCRIPTIVO O ESTRUCTURAL:

El marcado descriptivo, y por tanto, un **lenguaje de marcado descriptivo** (*Descriptive Markup Language*), o también denominado **lenguaje de marcado estructural** (*Structural Markup Language*), se basa, como decíamos, en las ideas anteriormente expuestas sobre la codificación genérica, esto es, pone mayor énfasis en la descripción del propósito del texto

de un documento que en su apariencia física, diferenciando, por tanto, lo que el texto *es* frente a cómo se *ve*. El concepto básico del marcado descriptivo se basa, pues, en la premisa de que el contenido de un documento debe permanecer separado de su estilo de presentación[44], requiriendo, por tanto, para esto último otro proceso distinto de formateado. De este modo es posible, al enfatizar la descripción de la estructura del documento electrónico, establecer múltiples formatos de presentación de una misma información (para ser presentada en la pantalla del ordenador, en papel, en soporte CD-ROM, en un espacio electrónico en línea como es Internet, etc.)

Este lenguaje utiliza las **marcas descriptivas** (*descriptive tags*) para cualificar cada objeto del documento, como primer paso en su transformación de información procesable por medios electrónicos: cada etiqueta descriptiva permite tratar una unidad de información como un objeto o entidad a la que se le pueden atribuir una serie de características específicas que pueden ser interpretadas y tratadas automáticamente por el ordenador. Los datos se transforman en objetos cualificados con una serie de atributos que los definen.

De esta definición de objetos o entidades lógicas dentro del texto electrónico mediante la utilización del marcado descriptivo se puede establecer o inferir, como señala Sperberg-McQueen, otra de las características esenciales, y ya anunciada, del marcado descriptivo: la estructura y las propiedades del texto[45]. La mayoría de los documentos que maneja el ser humano tienen establecidas una serie de convenciones o formalismos tipográficos procedentes de nuestra cultura impresa, lo cual, si se le añade el conocimiento lingüístico, gráfico y semiótico que poseemos sobre el lenguaje natural en el que están escritos los documentos, nos permite identificar de forma más o menos clara las partes o elementos que conforman su estructura. Según G. Salton, es posible diferenciar dos tipos de

---

[44] *SGML: Getting Started. Op. cit.* http://www.arbortext.com/Think_Tank/SGML_Resources/Getting_Started_with_SGML/getting_started_with_sgml.html

[45] C. M. Sperberg-McQueen, Claus Huitfeldt, Allen Renaer. *Meaning and Interpretation of Markup* [documento HTML]. Comunicación presentada al ALLC/ACH 2000 Conference (Glasgow, Scotland), 21-25 july, 2000. Disponible en http://www2.arts.gla.ac.uk/allcach2k/Programme/session2.html (consultado el 7 de agosto de 2000).

estructuras dentro de todo documento: una estructura abstracta y una estructura semántica[46].

La **estructura abstracta** es definida como aquella que especifica cómo se ajustan las diferentes piezas que conforman el documento, frente al concepto de **estructura semántica**, interesada en el significado del texto, tanto del conjunto como de cada una de las piezas que lo componen. La estructura abstracta se subdivide a su vez en otros dos tipos de estructuras: la estructura física y la estructura lógica.

La **estructura lógica** (*logical structure*) sería la definición de la composición de los objetos o piezas lógicas que se integran dentro de los grandes componentes que conforman los documentos electrónicos, así como el modo en el que dichos objetos se agrupan, se suceden y se relacionan entre sí dentro del documento. Normalmente las estructuras lógicas de los documentos electrónicos serán similares según tipos o familias de documentos, esto es, estructuras lógicas genéricas (por ejemplo, de forma general un libro puede estar dividido en capítulos, éstos en secciones, éstos en párrafos, etc., siguiendo un orden jerárquico establecido). La **estructura física** (*physical structure*)[47] describiría la composición y aspecto físico del documento.

---

[46] Gerard Salton, Chris Buckley, James Allan. "Automatic structuring of text files" [documento PDF]. *Electronic Publishing*, v. 5, nº 1, March 1992, p. 2. Disponible en http://cajun.cs.nott.ac.uk/compsci/epo/papers/volume5/issue1/ep056gs.pdf (consultado el 19 de septiembre de 2000).

[47] La estructura física es denominada en ocasiones por ciertos autores angloparlantes indistintamente como *layout structure* o, también, como *geometric structure*. Aunque en muchos casos todas estas denominaciones apuntan a una misma definición, esto es, a la estructura que define la presentación física o visual de los objetos contenidos en el documento, no es menos cierto que pueden existir sutiles diferencias entre todos estos términos. Así, la *physical structure* definiría, en esencia, cómo y dónde se almacenan y relacionan los objetos contenidos en un documento (pensemos en un documento multimedia del tipo HTML compuesto por ficheros distintos de texto, de imágenes, de audio, de video, etc.). La *layout structure* es la que habitualmente se toma para definir el concepto de estructura física del documento pues, en este caso habitual, se detalla la presentación visual de los objetos del documento. Por último, en el caso de la *geometric structure* el concepto es similar al anterior término, pero en este caso empleado en el campo científico de la digitalización de documentos, dado que definen los objetos del documento a través de las representaciones geométricas que se producen dentro del mismo al emplear estas técnicas. Sin duda, la mejor orientación a este respecto se pueda encontrar en la clasificación de modelos estructurales de documentos establecida por V. Quint al diferenciar dentro de la **estructura física** de los documentos entre la **macro-estructura**, correspondiente a la presentación y ocupación del espacio del documento por parte de las diferentes zonas que componen el área definida por el documento en su conjunto (correspondería a las denominadas *layout sctructure* y *geometric structure*) y la **micro-estructura**, correspondiente al soporte empleado para el documento y lugar de

Los lenguajes de marcado descriptivo, por tanto, pondrán un especial interés en la descripción de las estructuras lógicas que definen a los documentos electrónicos. De este modo, cuando la información contenida en los documentos es analizada y dividida, se definen sus partes y las relaciones que se establecen entre éstas, el procesamiento de dicha información por parte de las computadoras mejora y su utilización puede ser múltiple y variada. Con todo ello, es posible crear un modelo de datos que sea independiente de la plataforma y del software que lo genera, posibilitando de este modo, y entre otras cosas[48]:

- Que la información contenida pueda ser reutilizada de forma automática.
- Que la información pueda ser fácilmente compartida entre diferentes aplicaciones informáticas.
- Que la información pueda ser organizada en bases de datos para su posterior búsqueda y recuperación.

Tal es, por tanto, la importancia de la definición de los componentes estructurales del documento que algunos autores citan a este modelo descriptivo como lenguaje de marcado basado en la estructura (*structural markup language*) [49].

El principal problema en el lenguaje de marcado descriptivo estará, por tanto, en la identificación de los elementos que configuran la estructura de un documento, siendo esta elección el aspecto más delicado del análisis del documento electrónico, tal y como lo señala DeRose al afirmar que *the most point to be emphasized is that "how you divide up your data does matter"*[50].

---

almacenamiento de los diversos elementos que componen el documento (correspondería en este caso a la denominada con anterioridad *physical estructure*). Para una mayor información a este respecto, véase el documento de Vincent Quint. *Édition de documents structurés* [documento PostScript]. INRIA, 1994. Disponible en ftp://ftp.inrialpes.fr/pub/opera/rapports/CoursAix.ps.gz (consultado el 12 de septiembre de 2000).

[48] M. P. Box. *Op. cit.*, http://www.paradigma.com.br/XML/introxml.htm

[49] L. Alschuler. *Op. cit.*, p. 25.

Esta visión estructural del contenido del documento electrónico reporta innumerables beneficios para el procesamiento y gestión de dichos documentos, como lo resumen André, Furuta y Quint, al señalar que a partir de dicha definición del documento permite, por ejemplo, construir tablas de contenido o índices, numerar de forma automática secciones o notas, ordenar y componer el documento de múltiples maneras, establecer diversos estilos de presentación del documentos sin que haya que cambiar el texto del mismo, enviarlo y compartirlo con otras personas, etc[51].

Por todo lo dicho hasta el momento, los beneficios que se derivan de la utilización de los lenguajes de marcado descriptivo frente a los modelos basados en el marcado de texto orientado al formato, pueden ser resumidos en los siguientes[52]:

- **Independencia informática:** Se trata de modelos construidos desde la independencia de plataformas informáticas y sistemas operativos, por lo que se configuran como modelos abiertos con aspiraciones a ser convertidos en estándares de uso público.
- **Uniformidad en el formato:** reduce en gran medida el problema de la uniformidad de estilos o formatos de presentación cuando se trabaja en proyectos de elaboración documental integrados por múltiples autores, dado que lo que se marcan son las líneas generales y comunes referentes a la estructuración de los contenidos, no a la presentación física de los mismos (la cual será establecida en una fase posterior de trabajo y, en muchos casos, por un equipo distinto).

[50] Steven J. DeRose. "Navigation, Access, and Control Using Structures Information". *American Archivist*, v. 60, nº 3, Summer 1997, p. 299.

[51] J. André, R. Furuta, V. Quint. *Structured documents*. Cambridge: Cambridge University Press, 1989, p. 6. Citado en Paul Stiff. "Structuralists, stylists and forgotten readers". *Information Design Journal*, v. 7, nº 3, 1994, p. 230.

[52] Este listado de beneficios aportados por los lenguajes de marcado descriptivo o estructural ha sido confeccionado teniendo en cuenta las opiniones de diversos autores especialistas en la material, aunque de forma principal debemos destacar lo señalado al respecto en el artículo de Malcolm Clark. Structural defects: form and content in electronic publishing. *Information Design Journal*, v. 8, nº 2, 1996, p. 156.

- **División de tareas:** El marcado descriptivo establece una beneficiosa y natural separación entre contenidos y reglas de representación. Esto conlleva que se pueda establecer una división clara de trabajo y funciones en cuanto a las personas que se dedican a elaborar los contenidos informativos de los documentos electrónicos y aquellas otras dedicadas a la composición y diseño de estos documentos que han de ser procesados y presentados en uno o varios medios de difusión. Y. Marcoux y M. Sévigny denominan a este factor de racionalidad y división del trabajo como *work factorization* en la producción de documentos electrónicos[53].
- **Intencionalidad definida:** Este tipo de marcado de documentos asegura que las intenciones con las que un autor ha redactado un texto queden perfectamente definidas y detalladas, evitando cualquier tipo de ambigüedad. Se describen los elementos constituyentes del documento y la importancia que éstos tienen dentro del mismo.
- **Portabilidad**: de forma general, los documentos electrónicos marcados con estos lenguajes se codifican en lenguaje ASCII (*American Standard Code for Information Interchange*), el método más sencillo, seguro y universal para que estos documentos puedan ser transferidos de un ordenador a otro a través de redes informáticas.
- **Estabilidad y longevidad**: la independencia física y lógica que proporcionan estos lenguajes de marcado de documentos asegura una estabilidad del formato y, por tanto, una longevidad mayor de los documentos electrónicos. Esto es, los documentos electrónicos no se construyen teniendo en cuenta la aplicación informática que los va a procesar, tan sólo se etiquetan los elementos estructurales y sus contenidos informativos. Este hecho posibilita desarrollar nuevas aplicaciones informáticas para el procesamiento de estos documentos sin tener por ello que modificar el marcado de los mismos.
- **Reutilización del documento**: al identificar y diferenciar cada una de las partes esenciales que se integran en el documento electrónico, resulta relativamente sencillo para cualquier aplicación informática extraer alguna de estas partes y reutilizarlas para otros fines o propósitos. Normalmente esta reutilización de piezas documentales

[53] Y. Marcoux, M. Sévigny. *Op. cit.*, p. 588.

generará nuevas versiones de los documentos electrónicos existentes o, en muchos otros casos, nuevos documentos electrónicos de tipología distinta a la original, como diccionarios, glosarios, tesauros, boletines de resúmenes, etc.[54]

- **Búsqueda y recuperación documental:** el marcado descriptivo, basado en la adscripción de cada una de las partes del texto a un elemento estructural del documento electrónico (título, encabezamiento, párrafo, cita, pie de página, bibliografía, etc.), reporta innumerables ventajas, tal y como se verá posteriormente, a los sistemas automáticos de indización de documentos, así como a los sistemas de búsqueda y recuperación de información en medios electrónicos.

A modo de resumen, y utilizando las palabras del propio Goldfarb, un lenguaje de marcado descriptivo es aquel que describe la estructura y otros atributos de un documento independientemente del sistema informático utilizado y de cualquier procesador que pueda representarlo[55].

La aplicación del marcado descriptivo a los textos electrónicos tuvo como primera expresión el desarrollo a finales de los 60 y principio de los 70 del lenguaje GML (*Generalized Markup Language*) por parte de la compañía IBM, que, como veremos en el capítulo siguiente, es algo más que un simple lenguaje de marcas y dará origen al buque insignia de los lenguajes de marcado, el SGML (*Standard Generalized Markup Language*).

---

[54] David M. Levy. "Document reuse and document systems" [documento PDF]. *Electronic Publishing*, v. 6, nº 4, December 1993, p. 339. Disponible en http://cajun.cs.nott.ac.uk/compsci/epo/papers/volume6/issue4/ep6x4dml.pdf (consultado el 19 de septiembre de 2000).

[55] Ch. F. Goldfarb, Y. Rubinsky. *Op. cit.*, p. 137.

CAPÍTULO II.2

DE UN LENGUAJE DE MARCADO GENERALIZADO A UN METALENGUAJE: DEL GML AL SGML

II.2.1. INTRODUCCIÓN

Antes de describir la importancia que tuvo la aparición de los lenguajes de marcado generalizado y su posterior proceso evolutivo, resulta conveniente para su correcto entendimiento hacer una pequeña introducción sobre la estrecha relación existente entre la producción de documentos electrónicos y el procesamiento de su contenido informativo para la generación y alimentación de sistemas de bases de datos.

Es en la década de los sesenta cuando la llamada *explosión de la información* empieza a tomar verdadero cuerpo: son cada vez más las empresas e instituciones de todo tipo que incluyen sistemas informáticos para la edición de documentos electrónicos y, por extensión, sistemas capaces de tratar de forma automática la información contenida en éstos debido al alto valor político y económico que se le asigna a los recursos informativos, tanto externos como internos. Los sistemas de creación, almacenamiento, distribución y acceso a la información, que tradicionalmente se habían basado en el papel impreso como fuente exclusiva de transmisión de información, comienzan a desplazarse hacia entornos automatizados basados en el uso de ordenadores. La facilidad y rapidez en la creación de documentos electrónicos que proporcionan muchas de las herramientas informáticas de procesamiento de textos electrónicos hace que el número de documentos de este tipo crezca de forma desmedida en muchas organizaciones. Pero, en la mayor parte de los casos, la gestión de la información generada se seguía basando en el documento impreso. Tan sólo se había creado un entorno electrónico para agilizar la producción documental: de la máquina de escribir se pasará a los procesadores de texto.

Por otra parte, la incorporación por parte de las empresas e instituciones de los sistemas informáticos de generación de bases de datos para el almacenamiento y recuperación de información electrónica supusieron un gran paso para una correcta, ágil y eficaz gestión de la información interna generada, y/o de la información externa seleccionada. Pero en esta época las tecnologías para el procesamiento de texto electrónico y las tecnologías de las bases de datos textuales eran, en gran medida, incompatibles: por un lado se generaban documentos electrónicos en diversos formatos y soportes heterogéneos (documentos

generados por procesadores de texto de diversas compañías comerciales, hojas de cálculo, etc.) y, por el otro, se implementaban sistemas automatizados de generación de bases de datos documentales con sus propios formatos de fichero electrónico. En estos sistemas de bases de datos textuales, los documentos eran representados como listas de palabras sin tener en cuenta para nada que los documentos llevan en sí una estructura jerarquizada de elementos (como secciones y párrafos) en cierto modo semejante a la estructura de una base de datos[56]. Además, los documentos electrónicos basados exclusivamente en formatos de presentación propietarios no podían ser procesados directamente por las herramientas informáticas de gestión de bases de datos dado que en dichos documentos no se detallaba nada que hiciera referencia a los elementos estructurales del texto electrónico y al contenido informativo en ellos. Como acertadamente señalan J. V. Rodríguez y P. M. Díaz, los intercambios de información que se producían entre las diferentes aplicaciones informáticas puestas en juego requerían, las más de las veces, complejas conversiones entre formatos distintos y, a menudo, incompatibles[57].

Los documentos electrónicos y las bases de datos distaban bastante entre sí en cuanto a su estructuración y su funcionalidad. El proceso que se llevaba a cabo en muchas organizaciones es sencillo de entender:

1. Se generaban de forma relativamente fácil y rápida documentos electrónicos de diversa tipología en los que primaba la presentación impresa, basados en un determinado formato de codificación y marcado del texto.
2. Dichos documentos, ahora en formato impreso, sufrían un proceso tradicional de análisis documental para determinar los mejores puntos de acceso a la información de interés.

[56] Brian Lowe, Justin Zobel, Ron Sacks-Davis. "A Formal Model for Representation and Querying of Structured Documents". *Journal of System Integration*, v. 7, nº 1, 1997, p. 31.

[57] José Vicente Rodríguez Muñoz, Pedro Manuel Díaz Ortuño. "Arquitectura de la información: XML y Web". En: Maria Eulalia Fuentes i Pujol (dir.). *Anuari de Biblioteconomia, Documentació i Informació: BIBLIODOC 2000*. Barcelona: Col.legi Oficial de Bibliotecaris-Documentalistes de Catalunya, 2001, p. 153.

3. Esta información de interés, transformada ahora en datos, era nuevamente "tecleada" y almacenada en campos y registros de bases de datos con sus estructuras y formatos de fichero electrónico propios.
4. La interrogación de dichas bases de datos para la obtención de determinada información de interés para la organización podía dar como resultado la generación de nuevos documentos electrónicos, con lo que la rueda daba una nueva vuelta.

Como se puede apreciar fácilmente, todo este proceso de información documental resulta realmente complejo, sofisticado y, sobre todo, poco operativo. Como se señala en un *White Paper* de la compañía americana Arbortext, el que la información tenga que pasar por tantos estadios en este entorno de trabajo, esto es, ser recopilada, clasificada, organizada y ensamblada de forma manual dentro de sistemas informatizados de bases de datos, producía, al margen de la lentitud, frecuentes errores en cada una de estas etapas del proceso de gestión documental dentro de las organizaciones[58].

Por otro lado, si se optaba por desechar todo este proceso de análisis documental para la alimentación de las bases de datos y sustituirlo por sistemas de recuperación de información electrónica más o menos complejos basados en la búsqueda a texto completo, los resultados podían ser aún peores. Muchos de estos sistemas resultaban ineficaces cuando el volumen de documentos electrónicos a interrogar era muy elevado dado que las posibilidades de generar "ruido" en la búsqueda aumentaban de forma considerable.

Los documentos electrónicos, basados en gran medida en una larga tradición de siglos de producción documental impresa en los que primaba el estilo de presentación, veían constreñida su potencialidad como vehículo de transmisión de información al no estar claramente diferenciados aspectos tales como la definición de elementos estructurales, los datos que componen dicha información y el corsé que imponía un determinado estilo de presentación documental. Pero, como señala Goldfarb, este hecho no ocurría con las bases

[58] *SGML: Getting Started. Op. cit.*
http://www.arbortext.com/Think_Tank/SGML_Resources/Getting_Started_with_SGML/getting_started_with_sgml.html

de datos informatizadas, en las que por principio se establece una clara distinción entre los datos reales que constituyen el documento, los formularios de entrada de esos datos y los informes de presentación solicitados[59]. Esta forma de actuación conlleva una auténtica independencia de los datos; de la información, al fin y al cabo.

Este principio de independencia de los datos contenidos en los documentos electrónicos se basa en las ideas propuestas por la corriente de gestión de recursos empresariales denominada *Open Information Management* (OIM)[60]. La idea central reside en que la información debe estar abierta al procesamiento por parte de cualquier programa informático, independientemente del programa que generó esos datos. Este principio de independencia, según el OIM, establece que los datos deberían ser almacenados dentro de los ordenadores con representaciones que no fueran propias de un determinado programa informático sino, más bien normalizadas, incluso cuando estos datos constituyan el contenido en sí del documento.

Con esta nueva orientación en la definición de información en formato electrónico, el texto podría, pues, ser diseccionado en tres componentes claramente diferenciados: el contenido, la estructura y el estilo. El contenido sería la información propiamente dicha del documento, la estructura definiría cómo se organiza la información dentro del mismo y el estilo establecería como plasmar de forma visual el documento.

[59] Charles. F. Goldfarb, Paul Prescod. *Manual de XML*. Madrid: Prentice Hall Ibérica, 1999, p. V.

[60] *Ibid.*, pp. V-VII.

II.2.2. GML (GENERALIZED MARKUP LANGUAGE)

Basándose en todas estas ideas anteriormente expuestas, y partiendo de las aportaciones realizadas por los lenguajes de codificación genérica y el marcado descriptivo, Charles F. Goldfarb junto a una serie de investigadores de la compañía IBM desarrollaron a principios de la década de los 70 un lenguaje de marcado de texto electrónico denominado **GML** (*Generalized Markup Language*), o **Lenguaje de Marcado Generalizado.**

Charles F. Goldfarb, graduado en Derecho por la Universidad de Harvard en 1964, entró en noviembre de 1967 en la compañía IBM gracias a los consejos de un amigo que descubrió que la afición de Goldfarb, escribir instrucciones de ruta para las carreras de rally, era susceptible de ser automatizada a través de programación informática y, por tanto, de ser comercializada (aunque, como el propio Goldfarb reconoce, sabía bastante poco de computadoras y de programación informática)[61].

Su primer trabajo relacionado con los lenguajes de marcado de textos electrónicos consistió en la instalación de un sistema de tipografía electrónica para ser aplicado a un periódico local, siendo igualmente su primera experiencia con la gestión de una base de datos para tratar de forma automática los documentos de una organización. Aunque sus inicios no fueron especialmente brillantes dentro de la empresa (a punto estuvo de abandonar este trabajo y retornar a su actividad anterior, la abogacía)[62], en 1969 sería asignado por la compañía a un proyecto de investigación que cambiaría radicalmente su vida. Ese proyecto, desarrollado en el centro científico de Cambridge de IBM, trataba de desarrollar un modelo viable para integrar el uso de los ordenadores a la labor cotidiana en las oficinas judiciales. Su objetivo principal consistía en establecer un modo operativo de

[61] Charles F. Goldfarb. *The Roots of SGML – A Personal Recollection* [documento HTML]. SGMLsource, 1996. Disponible en http://www.sgmlsource.com/history/roots.htm (consultado el 26 de agosto de 2000).

[62] Florita Sheldon, Thomas L. Warren. *Introduction to SGML* [documento HTML]. Oklahoma: Computer Assisted Technology Transfer Research Program, Oklahoma State University, [sin fecha]. Disponible en

trabajo informatizado en línea que integrase las tareas de edición de textos electrónicos con un sistema de recuperación de información y un programa de composición de páginas en dicho entorno automatizado: los documentos deberían ser producidos con un procesador de texto para ser posteriormente almacenados en una base de datos, así como preparados para su presentación por parte del programa de composición.

Goldfarb vio que entre los diversos sistemas utilizados era imposible la comunicación y el intercambio de datos debido a que cada uno de ellos utilizaba diferentes modelos de marcado del texto. Con la ayuda de dos programadores de la empresa, trató de desarrollar un método que permitiese la completa integración de estos sistemas. En principio, Goldfarb estableció que para llevar a cabo este cometido era necesario cambiar por completo el marcado de procedimiento que utilizaban estos programas informáticos por otro de tipo genérico, tal y como lo habían establecido anteriormente William Tunnicliffe[63], Stanley Rice y Norman Scharpf.

En ese mismo año, Goldfarb junto con Edward J. Mosher y Ray Lorie crearon un nuevo lenguaje de marcado al que denominaron en un principio *Text Description Language* (TDL), el cual sería utilizado para su modelo de sistema de procesamiento integral de textos electrónicos en el ámbito jurídico (edición de texto, almacenamiento y recuperación documental y composición del texto). El modelo gustó tanto a la compañía IBM que decidió que podría ser aplicado a cualquier ámbito profesional en el que se requiriese el procesamiento de textos electrónicos. A este proyecto experimental del sistema integrado de textos electrónicos se le denominó *Integrated Text Processing*, y al primer prototipo aplicado se le bautizó con el nombre de *Integrated Textual Information Management Experiment* (INTIME). Las pruebas iniciales de este prototipo se llevaron nuevamente a cabo en el

http://www.okstate.edu/ind-engr/step/WEBFILES/Papers/SGML.html (consultado el 28 de agosto de 2000).

[63] La figura de William W. Tunnicliffe en el desarrollo de lenguajes de marcado de textos basados en el uso de marcas genéricas ha sido fundamental. Como reconoce el propio Goldfarb, sus numerosas contribuciones desde la GCA al desarrollo del GML y, posteriormente, al establecimiento y aceptación del SGML, serían de vital importancia. De hecho, gracias a sus esfuerzos y tesón la Marina de los Estados Unidos será una de las primeras instituciones de envergadura en adoptar el estándar SGML. Para una mayor información véase Harvey Bingham, Charles F. Goldfarb. *SGML: In memory of William W. Tunnicliffe* [documento HTML]. 12 y 19 de septiembre de 1996. Disponible en http://www.oasis-open.org/cover/tunnicliffe.html (consultado el 9 de septiembre de 2000).

Centro Científico IBM de la Universidad de Cambridge, Massachusetts, y desarrollado sobre un ordenador IBM System/360 modelo 67 y con sistema operativo CMS (*Cambrigde Monitor System*).

Este sistema fue presentado a la comunidad científica en el 33° congreso anual de la *American Society for Information Science* (Philadelphia, 15 de octubre de 1970), siendo publicado en el volumen 7 de la *ASIS Proceedings*[64]. La compañía IBM decidió que en esta comunicación no se hablaría nada del nuevo lenguaje creado, dado que ya intuían que podría resultar un producto potencialmente interesante y no era conveniente que la competencia se enterase de dicho desarrollo.

En 1971, Goldfarb establece definitivamente el nombre de **Generalized Markup Language** a este modelo de lenguaje de marcado de texto electrónico, basándose para ello en las iniciales de los apellidos de los tres investigadores que lo crearon (**G**oldfarb, **M**osher y **L**orie). Y lo denominó *marcado generalizado* debido a que, como es expresado por el propio Goldfarb, no restringía el procesamiento de los documentos electrónicos a una aplicación informática determinada, a un formateado de estilo o a un sistema operativo concreto; era generalizado en su propósito[65].

Finalmente, GML es dado a conocer bajo su propio nombre como implementación dentro del producto de la compañía para la gestión de textos electrónicos, denominado *Advanced Text Management System* (ATMS), en un trabajo publicado en mayo de 1973 con el título de *Design Considerations for Integrated Text Processing Systems*[66].

El éxito fue tal que la compañía IBM empieza a expandir este modelo del lenguaje basado en el marcado descriptivo al resto de sus programas de procesamiento de texto de ámbito

---

[64] Curiosamente esta comunicación del congreso mencionado no fue firmada por los tres autores que desarrollaron inicialmente este modelo, siendo sustituido R. Lorie por Theodore I. Peterson. Charles F. Goldfarb en un artículo publicado en 1997 por el *Journal of the American Society for Information Science*, conmemorativo y descriptivo de aquella comunicación original, explica cuáles fueron los motivos de este hecho. Para una mayor información véase Charles F. Goldfarb. SGML: The Reason Why and the First Published Hint. *Journal of the American Society for Information Science*, v. 48, n° 7, 1997, pp. 656-661.

[65] Ch. F. Goldfarb, Y. Rubinsky. *Op. cit.*, p. 7.

[66] Charles F. Goldfarb. *Design Considerations for Integrated Text Processing Systems.* IBM Cambridge Scientific Center Technical Report G320-2094, May 1973. Citado en Ch. F. Goldfarb. "The Roots of...". *Op. cit.*, http://www.sgmlsource.com/history/roots.htm

general. Y no sólo eso, gran parte del lenguaje GML empieza a ser implementado en los *mainframes* de las principales empresas del mundo de la publicación electrónica[67].

IBM, con Goldfarb a la cabeza de los desarrollos de mercado para los productos de edición e impresión de documentos en la delegación de San José (California), siguió aplicando y perfeccionando este modelo para tratar toda la vasta documentación que se generaba de forma electrónica dentro de la propia compañía, incluyendo la gestión de los documentos de todo tipo, desde manuales técnicos y notas de prensa hasta los contratos legales y la documentación de proyectos específicos. Esta apuesta permitió explotar y mejorar dicho lenguaje haciendo que éste evolucionase y surgiese otro formato más potente y robusto, el *Document Composition Facility Generalized Markup Language* (DCF GML).

En 1978 la compañía IBM publicó la Guía del Usuario del DCF GML[68], la cual incluía la primera descripción formal de un tipo de documento de propósito general. Este formato evolucionado de GML es el que utilizaban algunos de los productos comerciales de la casa IBM, como el *software* BookMaster y el BookManager[69]. DFC GML fue un producto comercial que se utilizó ampliamente en muchas empresas de todo el mundo dentro del campo de la edición y publicación de documentos electrónicos.

Después de la finalización del desarrollo de GML, Goldfarb continuó sus investigaciones acerca de las estructuras de los documentos, añadiendo conceptos adicionales a los ya desarrollados, como el de las referencias cortas, los procesos de hiperenlaces entre los documentos, y tipos de documentos concurrentes, que no eran parte de GML pero que sí se desarrollarían como elementos fundamentales dentro de SGML.

Dejemos de lado por un momento el importante proceso evolutivo del lenguaje GML para centrarnos en los aspectos más sobresalientes del mismo y las principales aportaciones realizadas en la construcción de un lenguaje de marcado de texto de propósito general, pues

---

[67] SGML Users' Group. *Op. cit.*, http://www.oasis-open.org/cover/sgmlhist0.html

[68] Documento interno de la compañía con número IBM SH20-9160.

muchas de ellas serán años más tarde contempladas por el estándar internacional SGML y sus posteriores lenguajes derivados.

Una de las características primeras y básicas que diferencian a GML frente a los anteriores lenguajes de marcado, incluidos aquellos de forma y propósito similares, como era el caso de GenCode, radica en su punto inicial de partida. Los anteriores lenguajes de marcado de texto se circunscribían al ámbito de la edición electrónica de documentos, en donde el fin primordial era la manipulación del documento para su presentación en un medio electrónico; no eran, en rigor, lenguajes de marcado descriptivo, tal y como se ha expuesto anteriormente aquí. GML se inscribe dentro de algo más global y cercano a los documentalistas, el procesamiento y gestión integral de la documentación electrónica. Es por este hecho tan diferenciador por lo que un lenguaje de marcado de estas características nada tuviera que ver, en un principio, con la presentación física de los documentos y sí con el intercambio y manipulación de datos entre diversos sistemas de computadoras para su almacenamiento y posterior recuperación en un entorno corporativo informatizado. Estos datos eran textuales, pues GML estaba orientado en un primer momento a este tipo de información, aunque con las posteriores mejoras, fue posible utilizarlo con imágenes digitales, gráficos, vídeos y sonidos.

Establecida esta primera e importante característica, y según palabras del propio Goldfarb citadas por M. Bryan, el desarrollo de GML (y su posterior versión DCF GML) se fundamentó en dos principios básicos[70]:

- El marcado de texto debería describir la estructura lógica del documento y otra serie de atributos frente a la apariencia o formato de presentación del documento.

---

[69] L. Alschuler. *Op. cit.*, p. 34.

[70] Ch. F. Goldfarb, E. J. Mosher, T. I. Peterson. "An Online System For Integrated Text Processing". *Proceedings of the American Society for Information Science*, n° 7, 1970, pp. 147-150. Documento citado en E. J. Martin Bryan. *SGML: An Author's Guide to the Standard Generalized Markup Language*. Reading: Massachusset: Addison-Wesley, 1998, p. 11.

- El marcado de texto debería evitar la ambigüedad y ser fácil de entender, tanto por las personas como por los programas informáticos.

Centrándonos en el primer principio, ya expusimos con anterioridad que todo lenguaje de marcado descriptivo conlleva, entre otras cosas, la identificación de la estructura lógica del documento sobre el que se está aplicando dicho marcado. En GML el autor identifica cada elemento significativo del documento y lo marca con aquel identificador genérico considerado como el mejor para caracterizar su función y naturaleza, al principio de dicho elemento y al final de éste. Para GML (y es algo que se impondrá posteriormente), el primer identificador genérico, el cual se sitúa al principio del elemento al cual va a caracterizar, estará delimitado por el símbolo de los dos puntos (:) delante y el símbolo del punto (.) detrás. El segundo identificador genérico, el cual se sitúa al final del elemento, estará delimitado por el signo de dos puntos seguido de la letra e (**:e**) delante y el símbolo del punto (.) detrás. La marca constituido por el nombre asignado al identificador genérico (normalmente, un nombre nemotécnico) y los dos signos que lo delimitan es lo que se conoce por etiqueta o *tag*. Cada pareja de etiquetas, la de inicio y la de fin, con su contenido, se denomina **elemento**. En algunos casos, la etiqueta final podía ser omitida dado que de este modo el usuario podía reducir el tedioso proceso de etiquetado (no ocasionando mayores inconvenientes para el sistema informatizado encargado de procesar este texto).

Un ejemplo de etiquetado con el lenguaje DDFC GML sería el siguiente[71]:

```
:h1.Chapter 1: Introduction
:p.GML supported hierarchical containers, such as
:ol.
:li.Ordered lists (like this one),
:li.Unordered lists, and
:li.Definition lists
:eol.
as well as simple structures.
:p.Markup minimization (later generalized and
```

[71] Ch. F. Goldfarb. "The Roots of...". *Op. cit.*, http://www.sgmlsource.com/history/roots.htm

```
formalized in SGML),allowed the end-tag to be omitted
for the "h1" and "p" elements.
```

Sin entrar en detalle en cuanto a las reglas de construcción y establecimiento de marcas en el lenguaje GML (pues todo ello se verá con mayor detalle en los apartados de los lenguajes que se derivaron de éste, SGML y, en la actualidad, XML), observamos que el documento ha sido marcado con etiquetas para identificar aquellos elementos del mismo que tiene una determinada función y naturaleza. Así, nos encontramos diferenciados aspectos tales como un encabezado del capítulo (:h1. reducción nemotécnica de *heading* de nivel 1), párrafos (:p. reducción de *paragraph*), una lista ordenada (:ol. y :eol. reducción de *ordered list*), con sus correspondientes ítem de lista (:li. reducción de *list item*), quedando perfectamente delimitado, con sus correspondientes etiquetas de inicio y de fin, el contenido textual de cada uno de estos elementos.

Además de esta identificación de piezas lógicas del documento, GML establece que la estructura es algo inherente en los documentos, sobre todo en aquellos que son generados y gestionados en organizaciones donde los procesos y las estructuras documentales están muy normalizadas. Si existen estructuras en los documentos, éstas pueden ser representadas como árboles virtuales en los que, consiguientemente, existe una organización jerárquica de los elementos que componen dicho árbol. De este modo, cada documento lleva asociada una determinada estructura lógica de sus elementos constituyentes, lo que le diferencia frente al resto de documentos que no pertenezcan a su misma clase o tipología. A esta importante propiedad introducida en GML se le conocerá posteriormente con el nombre de **Tipo de documento** (*document type*).

En cuanto al segundo principio mencionado, para que un lenguaje sea sencillo de entender debe basarse en una gramática sin reglas complejas ni códigos que no sean utilizados comúnmente por los humanos o por los programas informáticos. Pero sobre todo, un lenguaje de marcado debe evitar la ambigüedad, y esto se consigue con un marcado riguroso. Un marcado de este tipo implica que la definición de la jerarquía

estructural de los elementos, identificada simplemente por medio del marcado del comienzo y el final de cada elemento del documento. Por tanto, para GML no era necesario incluir información adicional para interpretar dicha estructura.

GML era, al fin y al cabo, como señala D. Connolly, una sencilla sintaxis para el marcado de documentos electrónicos, lo que permitía que muchos autores forzasen dicho lenguaje, omitiendo etiquetas que podrían parecer obvias, para crear un marcado minimizado que fuera sencillo de escribir y de leer por el ser humano[72]. Pero este modo de proceder, en un principio aceptable, era válido siempre que los tipos documentales a tratar fueran pocos; y así sucedía en aquellos momentos. No eran demasiados los tipos documentales que fueran requeridos a un mismo tiempo, por lo que los autores escribían compiladores especiales ajustados a cada forma particular de documento para manipular la codificación de los formatos de los datos de la forma más apropiada. A medida que fueron definiéndose más tipos documentales, con estructuras y características propias, se vio la necesidad de establecer un mecanismo de normalización para la generación y manipulación de cada uno de estos tipos documentales.

[72] Dan Connolly, Rohit Khare, Adam Rifkin. *The Evolution of Web Documents: The Ascent of XML* [documento HTML]. Pasadena: California Institute of Technology, Janaury 15, 1998. Disponible en http://www.cs.caltech.edu/~adam/papers/xml/ascent-of-xml.html (consultado el 17 de agosto de 2000).

II.2.3. SGML (STANDARD GENERALIZED MARKUP LANGUAGE)

En 1978 el Comité para el Procesamiento de la Información del *American National Standards Institute* (ANSI) crea el subcomité *Computer Languages for the Processing of Text*, el cual estará dirigido por Charles Card, de la compañía Univac, y que contará con la participación de Norman Scharpf, director de la *Graphic Communications Association* (GCA). Su objetivo final era el establecimiento de un estándar que normalizase las formas de especificar, definir y utilizar un lenguaje de marcado de documentos electrónicos. Este subcomité solicitó la integración en el mismo de Goldfarb para que aportase las nuevas ideas y últimos desarrollos que había realizado del GML, las cuales fueron apoyadas y relanzadas para la definición del estándar buscado[73].

El resultado de estos esfuerzos se concretó en 1980 con la publicación del primer borrador de trabajo (*working draft*) del nuevo lenguaje desarrollado, denominado *Standard Generalized Markup Language* (SGML). En 1983 la GCA recomendó el sexto borrador de trabajo de este lenguaje como el estándar de aplicación para la industria, siendo adoptado en esos momentos por diversas instituciones y organismos públicos del gobierno norteamericano.

En 1984, la *International Organization for Standardization* (ISO) empezó a trabajar con la ANSI para que dicho estándar tuviera un alcance internacional. Se crearon nuevos comités en cada una de estas instituciones para trabajar de forma paralela y conjunta en dicho desarrollo: la ISO creó el grupo de trabajo ISO/IEC JTC1/SC18/WG8[74], al frente del cual estaba James Mason de la US *Oak Ridge National Laboratory*, y la ANSI reestructuró su comité con el nombre de X3V1.8, estando al frente del mismo William Davis de la *SGML*

[73] SGML Users' Group. *Op. cit.*, http://www.oasis-open.org/cover/sgmlhist0.html

[74] El *Working Group* 8 del *Subcommittee* 18 del *Joint Technical Committee* 1 de la *International Organization for Standardization* y la *International Electrotechnical Commission.*

Associates. La dirección y coordinación de ambos grupos de trabajo recayó en Charles Goldfarb.

En 1985 apareció una primera propuesta de borrador para un estándar internacional de marcado de documentos electrónicos, siendo rápidamente adoptado por la Oficina de Publicaciones Oficiales de la Comunidad Europea. También en este año se creó el *International SGML Users' Group*[75] con Joan M. Smith al frente. Las aportaciones y experiencias de trabajo que realizaron estos dos grupos perfilaron el texto definitivo, procediéndose a su aprobación en 1986 (Norma ISO 8879:1986), publicándose en un tiempo récord, el 15 de octubre de 1986, debido a la gran expectación generada. Pero para ser exactos, la norma internacional incluye a SGML como un elemento dentro de un conjunto o paquete de mayores proporciones. De hecho, el nombre exacto del estándar internacional es ciertamente ambicioso, **Information Processing-Text and Office System-Standard Generalized Markup Language**[76], lo que da idea de esa doble función que debería de desempeñar: un lenguaje de marcado válido tanto para el procesamiento de textos electrónicos como de utilidad para la gestión de la información en entornos corporativos automatizados.

Un año después de su aparición, y con el conocimiento y opiniones aportadas por parte de los primeros usuarios, se procedió a una revisión de este estándar internacional. En realidad esta revisión tan sólo supuso la publicación el 1 de julio de 1989 de una corrección de los errores tipográficos detectados, la incorporación de algunas omisiones del texto original, así como la inclusión de notas aclarativas de uso para algunas partes del texto que habían quedado un tanto confusas o ambiguas. Posteriormente, el 19 de enero de 1990, el Comité SGML de la ISO documentó oficialmente estos errores en el WG8 N1035, siendo plasmados en la norma internacional como el Apéndice B[77].

---

[75] En la actualidad su nombre se ha extendido al de *International SGML/XML Users' Group*, siendo uno de las principales foros de debate y trabajo para el establecimiento de estos metalenguajes. Para una mayor información consúltese la información suministrada en su sito web, en http://www.isgmlug.org/

[76] International Organization for Standardization. *ISO 8879-1986 (E): Information Processing-Text and Office System Standard Generalized Markup Language*. Geneva: ISO, 1986.

[77] Ch. F. Goldfarb, Y. Rubinsky. *Op. cit.*, p. 594.

Desde su reconocimiento internacional, el estándar internacional SGML (ISO 8879), sigue en constante evolución, adaptándose a los nuevos tiempos y cambios tecnológicos. La labor es llevada a cabo por el ISO/IEC JTC1/SC34/WG1[78]. Todas estas adaptaciones han dado como fruto la inclusión de numerosos anexos en el texto original (por ejemplo, los anexos K y L para la adaptación del SGML al entorno de la Web), así como una gran cantidad de documentos de trabajo para futuras adaptaciones y revisiones de la norma internacional[79].

Pero, curiosamente, SGML no fue el primer estándar internacional de lenguaje para documentos electrónicos estructurados. De hecho en el mismo año de 1989 fue aprobado el estándar internacional ODA (*Open Document Architecture*, originalmente denominado *Office Document Architecture*) como norma ISO 8613[80]. Sin entrar en consideraciones sobre este estándar de características muy similares a SGML, sí resulta conveniente señalar que en aquellos momentos ODA constituyó un duro competidor del lenguaje SGML, aunque debido a sus menores prestaciones frente a SGML su utilización y, por tanto, su implantación internacional fue escasa[81].

Sin entrar en un exhaustivo detalle sobre el lenguaje de marcado de documentos electrónicos SGML, analizaremos a continuación las principales características que se derivan de su propio nombre, pues tal es la importancia y repercusión que han tenido los principios básicos en los que se sustenta SGML que muchas de las actuales aplicaciones en

---

[78] El *Working Group* 1 del *Subcommittee* 34 del *Joint Technical Committee* 1 de la *International Organization for Standardization* y la *International Electrotechnical Commission*

[79] Para una mayor información sobre los últimos trabajos que se han realizado o los que están en fase de revisión con respecto a SGML, recomendamos la consulta de Charles F. Goldfarb. *Project Editor's Review of ISO 8879* [documento HTML]. SGMLsource, 4 de marzo de 1999. Disponible en http://www.sgmlsource.com/8879rev/index.htm (consultado el 1 de septiembre de 2000)

[80] El capítulo 10 (*A comparison of SGML and ODA*, pp. 103-119) de la obra de Joan M. Smith describe ampliamente este estándar internacional y las diferencias existentes con respecto a SGML, por lo que sugerimos su lectura para una mayor profundidad sobre el tema. Joan M. Smith. *SGML and Related Standards: Document Description and Processing Languages.* New York [etc.]: Ellis Horwood, 1992.

el campo de la informática aplicada al procesamiento y gestión de documentos electrónicos se han basado en dichos principios.

STANDARD: la primera característica a destacar es, sin duda, el hecho mismo de que SGML es un estándar, y además de ámbito internacional. Que algo alcance la categoría de estándar oficial proporciona al producto una garantía de funcionalidad. Así es entendido por los usuarios del mismo, dado el lento y duro proceso de trabajo al que se ve sometido hasta alcanzar dicho estatus y, de igual modo, el gran número de organizaciones de todo tipo, nacionales e internacionales que se involucran en el desarrollo de un estándar oficial. Como se señala en un estudio del año 95 de la compañía Microstar Software Limited, la estandarización dentro del campo de las tecnologías de la información es la base de los sistemas abiertos, pues de este modo se asegura que la información electrónica pueda ser procesada e intercambiada con independencia del hardware, software y plataforma utilizada. De este modo, un gran número de productos informáticos de distintas compañías se adaptan para trabajar con este estándar, lo que redunda en una mayor confianza y utilización por parte de los consumidores finales. Estos usuarios serán libres de utilizar el producto informático y la plataforma de trabajo que deseen para la creación y el procesamiento de esos documentos electrónicos marcados mediante SGML debido a la independencia informática que proporciona un estándar. Esta independencia física y lógica reporta un enorme beneficio a usuarios particulares y corporativos a la hora de intercambiar información entre sus respectivos sistemas[82].

Por otro lado, que SGML sea un estándar internacional aplicado al campo de los formatos de documentos electrónicos deriva otra de sus principales aportaciones dentro de la gestión documental en entornos informatizados: la perdurabilidad o estabilidad de los

[81] Y. Marcoux, M. Sévigny. *Op. cit.*, p. 588.

[82] *An Investigation Into the Role of SGML In An Electronic Forms Environment: A Study Prepared for the Treasury Board Secretariat by Microstar Software Limited With Observations and Recommendations Formulated by the Project Advisory Group* [documento HTML]. OASIS, 31 de marzo de 1995. Disponible en http://www.oasis-open.org/cover/gift-ef-final.html (consultado el 25 de agosto de 2000).

documentos[83]. Cuando los documentos electrónicos son creados con un determinado formato propietario de una compañía informática, estos documentos se ven sometidos a las consideraciones que dicha compañía quiera establecer con su formato (cambios continuos de versiones, incompatibilidad con otros formatos propietarios de otras compañías, abandono de esa línea comercial del producto, posible desaparición de la compañía, etc.). Así, un documento electrónico creado según las normas SGML garantiza que si una determinada herramienta informática para la creación o gestión de este tipo de documentos se ve modificada o, incluso, si desaparece, los documentos permanecerán inmutables ante dichos cambios, siendo igualmente creados o gestionados por las nuevas versiones o por otras herramientas informáticas de otras compañías que trabajen con dicho lenguaje para el marcado de documentos electrónicos. Por tanto, la independencia de los datos que establece el marcado del documento mediante SGML asegura la transportabilidad de los datos del documento desde un entorno de *hardware* y *software* a otro distinto sin que se produzcan pérdidas de información.

Otro beneficio añadido que se deriva de este modelo estandarizado de codificación independiente de la plataforma y dispositivo de salida es que el documento, o partes del mismo, puede ser reutilizado para otras aplicaciones. Así, la división que proporciona la estructuración de los contenidos del documento permite que todas o algunas partes del mismo puedan ser extraídas y procesadas en otras aplicaciones para generar otros productos o servicios (por ejemplo, la extracción de las citas o el resumen contenido puede servir para generar boletines o listados de índices independientes)[84].

GENERALIZED: el término "generalizado" tiene múltiples significados o interpretaciones dentro de SGML. La noción más común que se tiene de este concepto, como ya se comentó anteriormente, es que los documentos electrónicos son marcados y almacenados con SGML de forma neutral, dado que los elementos informativos del

[83] Y. Marcoux, M. Sévigny. *Op. cit.*, p. 586.

[84] Stuart L. Weibel. "The World Wide Web and Emerging Internet Resource Discovery Standards for Scholarly Literature". *Library Trends*, v. 43, nº 4, spring 1995, p. 631.

documento son descritos a través de identificadores genéricos en lugar de identificadores de procesamiento específicos (por ejemplo, un párrafo puede venir precedido por la etiqueta <p> para identificarlo en lugar de un código concreto que especifique que dicho párrafo ha de ser procesado con un determinado tipo de letra y tamaño de la misma)[85]. Dado que el marcado del documento electrónico es generalizado, y no depende por tanto de ningún sistema específico, permite que sea tratado por cualquier herramienta informática de procesamiento de documentos, actual y futura. Por todo ello, y como señala acertadamente Marcoux, en esencia lo que es generalizado en SGML es el marcado, no el lenguaje en sí. De este modo, y como señala de forma humorística este autor, las siglas de este estándar internacional deberían ser analizadas como S((GM)L) y no como S(G(ML))[86]. Y esto es ciertamente así por el hecho de ser SGML algo más que un lenguaje, más bien un metalenguaje, tal y como se verá posteriormente. Así, si la utilización del marcado en SGML permite representar todos los tipos de documentos electrónicos posibles, no parece lógico que el estándar contemple todos los posibles identificadores genéricos para todos los tipos posibles. Es por este hecho, que al mecanismo que determina el conjunto de etiquetas a ser empleadas y las reglas a utilizar con dicho conjunto, este autor lo denomine como *generalized markup*.

MARKUP: el concepto de marcado de texto ya ha sido introducido anteriormente en este capítulo, al señalar que se trata de información añadida al documento electrónico para indicar algunas características de ciertas partes del mismo. El marcado en SGML está diseñado principalmente para identificar objetos dentro del documento de acuerdo a su función o propósito, y no, dado que es un lenguaje descriptivo, para detallar de las características tipográficas de dichos objetos[87].

---

[85] J. M. Smith. *Op. cit.*, p. 14.

[86] Y. Marcoux, M. Sévigny. *Op. cit.*, p. 586.

[87] Sharon C. Adler. "The Birth of a Standard". *Journal of the American Society for Information Science*, v. 43, nº 8, 1992, p. 556.

De forma general, la sintaxis concreta de SGML establece que los **elementos** empleados para marcar los objetos o piezas estructurales del documento electrónico harán uso de etiquetas de inicio en la forma de <IG>, y etiquetas finales, en la forma de </IG>. Los nombres de los elementos (o identificadores genéricos, IG, que normalmente serán sustantivos o adjetivos) que puedan aparecer en las correspondientes etiquetas, indican el tipo de información localizada entre las mismas. Los elementos identifican el marcado de las piezas lógicas que constituyen la estructura del documento a través de un identificador genérico (IG), o nombre de elemento, incluido en una etiqueta de inicio y otra final. Así, cada elemento encierra entre dichas etiquetas el contenido textual (o de cualquier otra índole) correspondiente a la pieza estructural que define. La etiqueta de inicio contiene dicho identificador genérico delimitado por delante con el símbolo de menor que (<) y por detrás con el símbolo mayor que (>), y la etiqueta final, diferenciada de la anterior por la inclusión de una barra inclinada entre el símbolo de menor que y el IG (</).

SGML implementa un mecanismo para reducir o simplificar el empleo de etiquetas en el marcado de los documentos electrónicos, permitiendo en ocasiones al usuario omitir algunas por quedar perfectamente delimitado su contenido con la utilización de una de las dos (caso muy habitual en los elementos considerados *vacíos*); por ejemplo, para explicitar que se está marcando una imagen o un gráfico tan sólo será necesario emplear la etiqueta de inicio, como se verá posteriormente). Además, las etiquetas de inicio pueden contener información añadida que matiza o detalla algún aspecto del contenido del elemento al cual está caracterizando, a través de los denominados atributos.

Los **atributos** de un elemento, cuando existen, son sustantivos o adjetivos que describen ciertas características del identificador genérico, y su valor viene definido por un conjunto de caracteres, o una referencia a una entidad, encerrado normalmente dentro de *delimitadores literales*, siendo estos representados por un par de comillas dobles (" ") o simples (' '). En cualquier caso, no está permitido el uso indistinto de ambos tipos de comillas dentro de los valores de atributos de una misma etiqueta.

Una **entidad** en SGML es el nombre de un objeto. La entidad, por tanto, puede ser una cadena de caracteres, un fichero completo de texto o un fichero binario, que ha de ser analizado por la aplicación informática correspondiente (*parser*) cuando en el documento

aparezca una referencia a dicha entidad. Se trata, pues, de un tipo de marcado que no es descriptivo, y sí de tipo referencial, tal y como se expuso al principio de este capítulo. Cada entidad externa o interna es referenciada en el documento SGML (*entity reference*) a través de un identificador con delimitadores distintos a los utilizados anteriormente: es precedido por el delimitador de apertura del *ampersand* (&), y seguido por el delimitador de cierre del punto y coma (;)[88]. Las entidades en SGML tienen, según palabras de Sperberg-McQueen, un especial significado dado que es posible incluirlas dentro de un documento marcado con SGML sin que ello conlleve consideraciones en torno a la estructura de dicho documento[89].

Por último, hay que señalar la utilización de **marcas de comentario** para incluir texto dentro del documento SGML que no ha de ser procesado por el programa, esto es, no será presentado en el dispositivo de salida ya que constituyen comentarios o anotaciones del autor que explican la utilización de ciertas marcas o indican el comienzo de ciertos bloques estructurales para la mejor comprensión de dicho marcado. Los comentarios van encerrados entre sus correspondientes etiqueta inicial, que en este caso es el signo menor que, el signo de cierre de admiración y dos guiones (<!--), y final, compuesta por dos guiones y el signo de mayor que (-->).

SGML fue diseñado originalmente para describir documentos de contenido textual, comúnmente codificados como ficheros de texto, en los que se pueden emplear los estándares de codificación EBCDIC o, principalmente, ASCII[90]. Esto permite, además de la facilidad para la conversión entre diferentes formatos si fuese necesario, que pueda ser utilizado cualquier procesador de texto existente en el mercado para realizar el marcado manual de los documentos electrónicos. Este hecho no quita que existan en el mercado

---

[88] En realidad existen dos tipos de entidades: entidades generales y entidades paramétricas. En el marcado existente en el documento SGML sólo se dan las entidades generales aquí definidas dado que las entidades paramétricas, tal y como se verá posteriormente, sólo se dan en la Definición de Tipo de Documento o DTD.

[89] "A Gentle Introduction to SGML" [documento HTML]. En: Sperberg-McQueen, Lou Burnard (eds.) *Guidelines for Electronic Text Encoding and Interchange (TEI P3).* Chicago: Academic Computing & Communications Center, University of Illinois, [sin fecha]. Disponible en http://www-tei.uic.edu/orgs/tei/sgml/teip3sg/index.html (consultado el 1 de septiembre de 2000).

[90] Pero, tal y como decíamos, tan sólo en principio pues SGML proporciona al mismo tiempo un mecanismo para apuntar o señalar datos no textuales, como gráficos, imágenes o sonidos, y datos en otros formatos predefinidos, tal y como se verá posteriormente en este capítulo.

editores especializados en SGML que automatizan este proceso de marcado, evitando que se produzcan errores en la inserción de marcas y, asimismo, ahorrando una considerable cantidad de tiempo al usuario en esta tarea.

LANGUAGE: de forma general, Marcoux define a SGML como un lenguaje informático de descripción de documentos electrónicos. Como todo lenguaje informático, SGML poseerá una sintaxis y una semántica: la sintaxis está basada en el uso de la codificación ASCII del texto electrónico y la semántica establece representaciones jerárquicas arborescentes de los elementos lógicos que componen dicho texto[91]. Pero para ser completamente exactos es necesario decir que, tal y como fue anunciado con anterioridad, SGML no es en sí un lenguaje de marcado de documentos electrónicos sino, más bien, un **metalenguaje**; esto es, un lenguaje que permite crear múltiples lenguajes de marcado descriptivo. Como señala D. Barron SGML es un metalenguaje que define, a diferencia de lo señalado por Marcoux, exclusivamente la sintaxis para un lenguaje de marcado generalizado y normalizado, ya que establece cómo debemos detallar el marcado de documentos, la sintaxis, pero no lo que significa este marcado, la semántica[92]. Por tanto, SGML define varias sintaxis para un lenguaje de marcado, proporcionando un mecanismo normalizado para la generación de una familia de lenguajes de marcado descriptivo que pueden ser utilizados, principalmente, para la descripción de la estructura de múltiples tipos de documentos electrónicos. Estos dos tipos de sintaxis son: una sintaxis abstracta, utilizada para declarar las reglas que definen la inserción de marcas descriptivas en los documentos sin tener en cuenta los caracteres específicos utilizados para representar dicho marcado; y otra sintaxis concreta, que detalla aspecto tales como los caracteres que van a

[91] Yves Marcoux. *Pourquoi SGML? Pourquoi maintenant?* [documento HTML]. Montreal: Faculté de Droit, Centre de Recherche en Droit Public, Université de Montréal, 8 juillet 1996. Disponible en http://www.lexum.umontreal.ca/fr/equipes/technologie/conferences/sgmlquebec/12.htm (consultado el 17 de noviembre de 2000)

[92] David Barron. "Why use SGML?" [documento PDF]. *Electronic Publishing*, v. 2, nº 1, April 1989, p. 8. Disponible en http://cajun.cs.nott.ac.uk/compsci/epo/papers/volume2/issue1/epdxb021.pdf (consultado el 19 de septiembre de 2000).

ser utilizados para los delimitadores de marcado, las cantidades, los nombres utilizados en las declaraciones de marcado, etc[93].

En el ámbito de los lenguajes de marcado a estos metalenguajes se les conoce por el nombre de **Generalized Markup Rule-Set** (GMRS) y, por tanto, SGML es en esencia un extenso y complejo GMRS con grandes capacidades para la definición de estructuras y nombres de elementos que han de ser declarados cuando se redacte un lenguaje de marcado generalizado concreto[94]. Así, uno de los lenguajes de marcado de documentos electrónicos más populares hoy en día, el *HyperText Markup Language* (HTML) no es otra cosa que la aplicación del metalenguaje SGML para la definición de un tipo de documento de propósito general y de aplicación en el entorno Web de Internet. Los GMRS o metalenguajes, y por tanto SGML, surgen ante la imposibilidad, como apuntábamos anteriormente, de establecer identificadores genéricos estandarizados para todos los tipos de documentos que pudieran existir. De este modo resulta más factible y operativo establecer unas reglas sintácticas para la definición de conjuntos de etiquetas[95]. Por tanto, la principal característica de los metalenguajes aplicados a este campo de actuación será su extensibilidad o ampliabilidad.

Los GMRS permiten que los autores puedan definir nuevos conjuntos de elementos descriptivos para detallar con exactitud cada uno de los tipos de documentos electrónicos que se desea tratar. Estos conjuntos de elementos pueden ser ampliados, modificados o, simplemente, eliminados. Así, cada conjunto fijo de elementos conforma un lenguaje de marcado generalizado y, por tanto, no conllevan esta cualidad de la extensibilidad. De hecho, cuando un lenguaje de marcado generalizado es modificado con la adición de nuevos elementos que nos permite la extensibilidad de los GMRSs nos encontramos ante un nuevo lenguaje generado, a pesar de que comúnmente se hable de nuevas versiones (por ejemplo, el lenguaje de marcado HTML 4.0 derivado del metalenguaje SGML es un nuevo

---

[93] Ch. F. Goldfarb, Y. Rubinsky. *Op. cit.*, pp.135-136.

[94] Richard Lander. *Introduction to Generalized Markup Language* [documento HTML] Waterloo, Ontario: University of Waterloo, [sin fecha]. Disponible en http://pdbeam.uwaterloo.ca/~rlander/XML/intro_gml.html (consultado el 23 de junio de 2000).

lenguaje con respecto a su versión anterior, el HTML 3.2). Este hecho no excluye que exista una "compatibilidad hacia atrás" en cada uno de estos nuevos desarrollos, lo que garantiza que cualquier sistema informático que sea capaz de procesar la última versión de cada desarrollo debe ser capaz, asimismo, de procesar las anteriores versiones.

II.2.3.1. BENEFICIOS APORTADOS

Si el propósito principal de cualquier documento es el de registrar, comunicar y compartir la información contenida, resulta innegable la importancia que tiene la búsqueda de un sistema capaz de localizar fácil y rápidamente la información deseada por el usuario, así como la validación y posible reutilización de dichos contenidos informativos. En este contexto, SGML se beneficia de las aportaciones realizadas hasta ese momento por las diversas propuestas y modelos de lenguajes de estructuración de los documentos electrónicos y de forma muy significativa, de los lenguajes de marcado orientados a la descripción de estructuras y contenidos.

La estructuración de los documentos electrónicos resulta pues un factor clave y determinante en el modelo SGML pues con ello se obtienen las ventajas derivadas de los modelos basados en la estructuración de la información (accesibilidad a partes concretas del documento, validación de los contenidos sujetos a una determinada estructura y la reutilización de contenidos del documento para múltiples propósitos, entre otras), añadiéndose todas aquellas que se derivan de los formatos estandarizados para el intercambio de textos a través de redes informáticas (formato independiente para el

[95] Y. Marcoux, M. Sévigny. *Op. cit.*, p. 586.

intercambio de datos electrónicos entre las organizaciones, longevidad del formato y respaldo por parte de la industria, entre otras)[96].

Con estas ideas de base se pueden destacar brevemente los beneficios que el estándar SGML puede aportar a cualquier sistema de procesamiento y gestión de documentos electrónicos, haciendo uso para ello de los puntos destacados por J. M. Smith en su manual; a saber[97]:

- Si la información de base del sistema de información es tratada mediante el marcado SGML estaremos ante un sistema riguroso, dado que SGML está basado en reglas estrictas que garantizan la integridad de la información.
- Un sistema documental basado en SGML conlleva una total integración entre todos los tipos de información existentes en la organización, tanto actuales como futuros.
- La división de estructura y formato que establece SGML permite que los documentos estén siempre al día, eliminando de este modo la necesidad de establecer copias o duplicados, con el riesgo que ello supone.
- Los sistemas SGML permiten que la información sea puesta a disposición de sus consumidores en múltiples y variadas formas, desde los medios impresos hasta modernos sistemas de hipertextos e hipermedios.
- Todos los elementos gráficos que pudieran existir dentro de un documento pueden ser fácilmente recuperados debido a la asociación que se establece de forma textual entre el fichero electrónico que contiene la imagen y los valores textuales establecidos para la descripción de la misma.
- Todos los elementos contenidos en el documento electrónico son inequívocamente identificables, con las ventajas que ello conlleva para el establecimiento de referencias cruzadas o la acotación de búsquedas documentales a un determinado elemento del documento.

[96] *Which is for Me? Structured Documents or SGML?* [documento HTML]. SGML Associates, 1996. Disponible en http://www.mcs.net/~dken/struct.htm (consultado el 5 de noviembre de 2000).

- La información marcada mediante SGML puede ser utilizada para múltiples propósitos que van desde la construcción de complejos sistemas de gestión de la información hasta potentes sistemas de publicación.
- Aunque los costes de desarrollo de sistemas SGML pueden ser elevados, la efectividad de los mismos está garantizada si se compara con otros sistemas, igualmente costosos, basados en lenguajes de formato propietarios.
- Y, por último, el hecho de que SGML sea un estándar internacional garantiza entre otras diferentes cosas que la información pueda ser intercambiada entre otros sistemas y la existencia en el mercado de un gran abanico de productos informáticos capaces de trabajar con estos documentos sin problemas de compatibilidad.

II.2.3.2 ESTÁNDARES RELACIONADOS

Desde la aprobación de la norma internacional SGML en 1986, la ISO ha venido trabajando en la elaboración de otros estándares relacionados, o también llamados "acompañantes", que complementan y potencian los documentos electrónicos SGML con nuevas capacidades para el intercambio, para la presentación y para la aplicación de las posibilidades de los hipertextos de dichos documentos en un entorno de trabajo informatizado. Este enriquecimiento del SGML a través de sus estándares acompañantes permite el desarrollo de un conjunto de facilidades generales que pueden ser compartidas entre todos esos estándares y que, cuando se integran dentro de los productos informáticos, permiten una notable facilidad de trabajo con los documentos electrónicos y una mayor potencialidad en el intercambio y suministro de información en un entorno

[97] J. M. Smith. *Op. cit.*, p. 51.

electrónico de trabajo[98]. Aunque no profundizaremos en detalle en este numeroso grupo de estándares acompañantes (pues nuestro principal interés se centrará en XML y sus propios estándares acompañantes), sí resulta conveniente describir de forma sucinta la función y algunas de las características más sobresalientes de tres de ellos debido a la gran importancia que han tenido. Éstos son:

- **SDIF (SGML Document Interchange Format) – ISO 9069:1988**[99]: Es el estándar internacional desarrollado para el intercambio de documentos electrónicos SGML. Debido a que la norma internacional SGML no detallaba cómo organizar cada una de las partes que componen el documento SGML (la DTD, las entidades, documentos y datos externos) para el intercambio de este tipo de documentos en un entorno informatizado de trabajo dominado, en aquellos momentos, por los protocolos OSI (*Open Systems Interconnection*) para la comunicación entre ordenadores a través de redes telemáticas[100]. El estándar SDIF proporciona, por tanto, un mecanismo normalizado y abierto para el empaquetamiento de cada una de las partes asociadas al documento SGML dentro de un flujo de datos en entornos telemáticos de comunicación, asociando para ello una serie de descriptores a cada una de esas partes para establecer las relaciones existentes entre todas ellas (por ejemplo, si un documento contiene la declaración de una entidad externa, ésta llevará asociada un descriptor en el flujo de datos, enlazando la declaración con dicha entidad). De igual modo, SDIF proporciona un mecanismo para que el receptor de ese flujo de datos pueda desempaquetar el contenido y determinar a través de los descriptores qué partes eran externas al documento antes de que el documento fuese empaquetado y poder, de este modo, reconstruir el documento con la configuración original.

---

[98] James David Mason. "SGML and Related Standards: New Directions as the Second Decade Begins". *Journal of the American Society for Information Science*, v. 48, nº 7, 1997, p. 593.

[99] International Organization for Standardization. *ISO 9069-1988 (E): Information processing – SGML support facilities – SGML Document Interchange Format (SDIF)*. Geneva: ISO, 1988.

- **HyTime (Hypermedia/Time-Based Structuring Language) – ISO/IEC 10744:1992**[101]: HyTime posibilita describir ciertos rasgos de los documentos SGML multimedia (texto, imágenes, gráficos, ficheros de sonido, etc.) y establecer relaciones entre las diferentes partes del documento multimedia a través de la creación de hiperenlaces entre dichas partes, pero sin definir el significado de estos rasgos o las relaciones que se establecen. HyTime debe su origen a tres corrientes distintas de trabajo, pero muy relacionadas en sus principios generales: al desarrollo de la aplicación informática DynaText, la cual generaba ficheros SGML dentro de un modelo de funcionamiento de hipertexto, a la creación por parte de Tim Berners-Lee, investigador del CERN (actual *Organisation Européenne pour la Recherche Nucléaire*), de una sencilla aplicación SGML para la generación de documentos genéricos con capacidades hipertextuales, y, en mayor medida, a la aplicación de SGML al medio musical a principios de los años 90 en la búsqueda de un estándar para definir piezas musicales. Este desarrollo, conocido como SMDL (*Standard Music Description Language*), incluía ciertas partes dedicadas a la inserción de características hipertextuales para la relación entre los distintos objetos que componían un documento SMDL. Debido a la importancia que fueron adquiriendo las partes dedicadas al hipertexto, la ISO estimó conveniente segregarlas de este estándar y comenzar un nuevo proyecto de forma independiente, denominado HyTime[102]. Así, HyTime es el estándar internacional ISO para la representación estructurada de documentos hipermedia e información basada en la sincronización temporal de diversos eventos incluidos en el documento electrónico (audio, vídeo, música, etc.), junto a las capacidades de los hiperenlaces para vincular objetos externos al documento. HyTime utiliza la sintaxis SGML para representar estos enlaces hipertextuales, por lo que se puede considerar a este desarrollo como una aplicación directa de SGML. En 1997 apareció la segunda y definitiva versión de

---

[100] J. M. Smith. *Op. cit.*, p. 65.

[101] International Organization for Standardization. *ISO 10744-1992 (E): Information technology – Hypermedia/Time-based Structuring Language.* Geneva: ISO, 1992.

[102] E. v. Herwijnen. *Op. cit.*, p. 236.

HyTime (**ISO/IEC 10744:1997**), en la cual se añaden ciertos aspectos de interés, como es el caso del Anexo C para la normalización del formalismo de una meta-DTD (conjunto de meta tipos de elementos), o tipo de plantilla denominada *architectural form* o *enabling architecture*[103].

- **DSSSL (Document Style Semantics and Specification Language) – ISO/IEC 10179:1996**[104]: DSSSL define la semántica y la sintaxis de un lenguaje específico para la definición del procesamiento de documentos SGML con vistas a su presentación en diferentes medios. Se trata de uno de los estándares internacionales en el que más se ha trabajado pues su comité de desarrollo se estableció en 1988 pero no se publicó definitivamente como norma internacional hasta enero de 1996. En sí, DSSSL proporciona cuatro áreas distintas de normalización[105]: un lenguaje y un modelo de procesamiento para la transformación de un documento SGML en otro documento SGML distinto o, en muchos otros casos, a un formato de documento electrónico distinto[106]; un lenguaje para especificar a la aplicación informática de procesamiento las características de formato para un determinado documento SGML; un lenguaje de interrogación denominado SDQL (*Standard Document Query Language*); y, por último, un determinado lenguaje de expresión, que forma parte de los dos anteriores lenguajes, utilizado para la creación y manipulación de objetos. El modelo conceptual de DSSSL,

[103] Una buena descripción general sobre arquitecturas SGML a través de HyTime se encuentra disponible en Steven R. Newcomb. *SGML Architectures: Implications and Opportunities for Industry* [documento HTML]. OASIS, [sin fecha]. Disponible en http://www.oasis-open.org/cover/newcomb-sgmlarch.html (consultado el 17 de diciembre de 2000).

[104] International Organization for Standardization. *ISO 10179-1992 (E): Information technology – Text and office systems – Document Style Semantics and Specification Language (DSSSL).* Geneva: ISO, 1996.

[105] Sharon C. Adler. "The ABCs of DSSSL". *Journal of the American Society for Information Science*, v. 48, nº 7, 1997, p. 597.

[106] Aspecto este último de vital importancia para muchas organizaciones dada la necesidad en ciertas ocasiones de transformar los datos SGML a un formato de documento electrónico distinto, por ejemplo, al HTML para la publicación de información en la WWW. Para una mayor información sobre estos procesos de transformación, véase Jon Fausey, Keith Shafer. "All My Data Is in SGML. Now What?". *Journal of American Society for Information Science*, v. 48, nº 7, 1997, pp. 638-643.

por tanto, contempla dos procesos distintos: un proceso de transformación de documentos SGML y un proceso de formateado de dichos documentos. Estos dos procesos pueden ser utilizados de forma conjunta o individualizada. El primero de estos dos procesos, el de transformación, proporciona etapas adicionales de procesamiento de documentos electrónicos que tradicionalmente no habían sido tratadas por otros estándares anteriores más orientados a la presentación, como era el caso de FOSI (*Format Output Specification Instance*). Estas nuevas etapas de procesamiento incluyen aspectos tales como la fusión o división de documentos, la generación de índices y tablas de contenidos, la adición de una estructura a la instancia de documento, la extracción de datos, y una validación adicional. Dentro de las capacidades de formateado soportadas por el estándar DSSSL se incluyen capacidades para el reconocimiento del contexto del objeto, para el reconocimiento de contenido específico dentro del documento SGML y una estructuración del formato de salida o presentación basado en el contenido[107].

Existen, como ya se ha comentado anteriormente, otros muchos estándares internacionales relacionados de una forma u otra con SGML (en especial dentro del grupo de las denominadas *SGML Support Facilities*, dentro de las cuales se incluye el estándar SDIF, descrito en primer lugar), lo que conforma una gran familia de normas y especificaciones subsidiarias y/o complementarias en continuo crecimiento[108].

Las primeras aplicaciones que se derivaron del uso del estándar internacional SGML fueron llevadas a cabo en pequeñas instituciones o por una comunidad reducida de usuarios. Sin embargo, este estándar empezó a ser adoptado rápidamente por grandes instituciones, como fue el caso de la *Association of American Publishers* (AAP), con la aplicación conocida bajo el nombre de *Electronic Manuscript Project*, o el Departamento de

[107] M. Colby. *Op. cit.*, p. 420.

[108] Eduardo Peis, Félix de Moya. "Sgml y servicios de información". *El profesional de la información*, v. 9, ° 6, junio 2000, p. 5.

Defensa de los Estados Unidos, para elaborar la documentación de la iniciativa *Computer-aided Acquisition and Logistic Support* (CALS). Son innumerables las organizaciones e instituciones públicas y privadas de múltiples campos de actuación que rápidamente empezaron a adoptar SGML como modelo de estructuración de contenidos documentales para la organización y gestión de la información propia, desarrollando aplicaciones que han tenido gran alcance y repercusión en todo el mundo. Baste destacar algunos de los principales esfuerzos en este campo como, por ejemplo[109], la *Electronic Publishing Special Interest Group* (EPSIG) y la *Graphic Communications Associations* (GCA) en el campo de la edición y publicación electrónica, el *Center for Electronic Texts in the Humanities* (CETH) con el conocido desarrollo TEI (*Text Encoding Initiative*) para el tratamiento y marcado de textos electrónicos de carácter humanístico, principalmente[110]; la *Air Transport Association* (ATA)

[109] Robin Cover mantiene desde mediados de la década pasada un magnífico recurso en la Web sobre los principales desarrollos que han ido surgiendo en torno a SGML. Véase en http://www.oasis-open.org/cover/sgml-xml.html

[110] Uno de los lenguajes de marcado descriptivo derivados de SGML que mayor repercusión ha tenido en todo el mundo por su aceptación y uso ha sido el lenguaje TEI (*Text Encoding Initiative*)[110]. Se trata de una iniciativa puesta en marcha a finales de los 80 por la *Association for Computers and the Humanities*, la *Association for Computational Linguistics* y la *Association for Literary and Linguistic Humanities* y respaldada económicamente por otras prestigiosas instituciones, entre las que destacan la *U.S. National Endowment*, la DG XIII de la Comisión de las Comunidades Europeas y la *Social Science and Humanities Research Council* de Canadá. Este lenguaje se orienta a la descripción estructural y semántica de textos humanísticos almacenados en formato electrónico y cuyo contenido, principalmente de carácter textual (aunque también se implementan mecanismos para la descripción de las imágenes que pudieran estar asociadas al mismo), ha de ser intercambiado entre instituciones e investigadores de todo el mundo a través de redes informáticas de comunicación. Dado el amplio espectro de tipos documentales que ha de cubrir (desde la traslación a formato electrónico de manuscritos medievales hasta obras literarias contemporáneas en verso o en prosa), este lenguaje se configura como amplio vocabulario de elementos y atributos de propósito múltiple, en el que han venido destacando por su influencia en otros desarrollos futuros de descripción de metadatos, los elementos de la cabecera de los documentos TEI (*TEI Header*). Dentro de esta cabecera se incluye toda aquella información necesaria para una correcta definición bibliográfica del documento. La definición formal (*guidelines*) o esquema de este lenguaje de marcado, conocida como TEI P3, fueron publicadas en mayo de 1994, existiendo una reimpresión revisada en mayo de 1999. Debido a la amplitud y la complejidad de uso de este lenguaje de marcado descriptivo se vio necesario hacer una adaptación reducida del mismo y que, además, estuviese redactada de forma más amigable para el usuario final, a modo de manual de uso con numerosos ejemplos. Esta reducción o subconjunto del esquema de codificación TEI P3 fue publicado en junio de 1995, y se le conoce internacionalmente bajo el nombre de TEI Lite (oficialmente denominado TEI U5). Desde la publicación de este estándar de facto para el marcado de textos electrónicos de contenido humanístico (principalmente, textos de contenido lingüístico y literario) han sido ciertamente numerosos los proyectos que se han desarrollado haciendo uso de sus directrices, la mayoría de ellos al amparo de instituciones universitarias de todo el mundo. Así, son de destacar dos proyectos de especial relevancia, como son *The Oxford Text Archive* (información disponible en http://ota.ahds.ac.uk/), proyecto de ya larga historia puesto

generadora de un gran número de normalizaciones de tipos documentales para manuales, documentación técnica, guías y glosarios electrónicos de aplicación en el mundo de la aviación; la *Telecommunications Industry Forum* (TCIF) con el desarrollo de TIM (*Telecommunications Industry Markup*), los esfuerzos de la *International Press Telecommunications Council* y la *Newspaper Association of America* en el desarrollo de una definición de tipo documental válida para los servicios de noticias, prensa y archivos en un entorno electrónico, conocida por el nombre de UTF (*Universal Text Format*); o, por último, la famosa y ampliamente utilizada aplicación DocBook para la descripción de manuales técnicos, principalmente de *hardware* y *software*, y mantenida actualmente por la *Organization for the Advancement of Structured Information Standards* (OASIS)[111].

De igual modo, la ISO estableció en 1994, con rango de estándar internacional, una serie de DTD normalizadas para cuatro tipos de documentos habituales en cualquier organización agrupadas bajo un único conjunto denominado *Information and Documentation – Electronic manuscript preparation and markup* (ISO 12083:1994)[112]. Estas cuatro DTD de propósito general definen la estructura y elementos para la edición electrónica de libros (ISO 12083:1994//DTD Book//EN), publicaciones seriadas (periódicos, revistas, etc.)

---

en marcha por los servicios informáticos de la Universidad de Oxford con el fin de proporcionar a la comunidad universitaria los textos en formato electrónico de un gran número de obras literarias clásicas en diversos idiomas (en la actualidad cuenta con más de 25.000 obras de 25 lenguas diferentes), y el *Center for Electronic Text in the Humanities* (CETH) (información disponible en http://www.ceth.rutgers.edu/), proyecto conjunto de las universidades americanas de Rutgers y Princeton, de similares características al anterior, donde, además, estos textos electrónicos se encuentran disponibles en la actualidad en formato HTML y XML. Este lenguaje de marcado se ha empleado de igual modo para el tratamiento de documentos jurídicos, siendo el caso más significativo el proyecto llevado a cabo por el *Center for Electronic Text in the Law* de la Universidad norteamericana de Cincinnati, del cual se hablará con mayor detenimiento en posteriores apartados de este capítulo. En cualquier caso, toda la información oficial relativa al lenguaje TEI se encuentra accesible en la dirección http://www.tei-c.org/

[111] Para una mayor información sobre esta aplicación SGML y la poderosa organización que lo sustenta véase http://www.oasis-open.org/docbook/

[112] La *Association of American Publishers* (AAP) desarrolló la versión original de estas DTDs a mediados de la década de los 80, las cuales serían normalizadas por la ANSI en 1988. Con ciertas modificaciones, estas DTDs se convertirían en un estándar internacional a través de la ISO en 1994. Desde entonces la ISO es la encargada de mantener este conjunto de definiciones de tipos documentales SGML a través de la *Electronic Publishing Special Interest Group* (EPSIG), institución ésta que incluye a la AAP y a la *Graphic Communications Association Research Institute* (GCARI) junto a otros miembros. Para una mayor información sobre estas DTDs de la ISO, véase International Organization for Standardization. *ISO 12083-1994 (E): Information and documentation – Electronic manuscript preparation and markup*. Geneva: ISO, 1994.

(ISO 12083:1994//DTD Serial//EN), artículos individuales (ISO 12083:1994//DTD Article//EN) e inclusión de fórmulas matemáticas en cualquier documento SGML (ISO 12083:1994//DTD Mathematics//EN). Pero sin lugar a dudas, la aplicación SGML que mayor repercusión ha tenido es HTML (*HyperText Markup Language*), tal y como se verá en la introducción del capítulo siguiente.

En el campo de la Biblioteconomía las experiencias y proyectos han sido múltiples y variados para el procesamiento, control y distribución de documentos electrónicos en las llamadas bibliotecas electrónicas o digitales[113]. Destacan proyectos en los que se ha venido aplicando el anteriormente citado desarrollo TEI (*Text Encoding Initiative*) a este entorno, o el caso de ELSA, DECOMATE y ELVYN orientados todos ellos al acceso y suministro de documentos electrónicos, utilizando SGML para su manipulación, procesamiento y transformación en sistemas de búsqueda y recuperación documental a texto completo o en sistemas de navegación hipertextual[114]. Incluso para algo tan cercano a los bibliotecarios y los documentalistas como es la descripción catalográfica de los documentos impresos, han sido numerosas las experiencias de aplicación del lenguaje SGML[115].

Las aplicaciones del lenguaje SGML a la documentación jurídica han sido también ciertamente numerosas e importantes, algunas de las cuales serán expuestas con cierto detenimiento en el último punto de este capítulo.

---

[113] Timothy W. Cole, Michelle M. Kazmer. "SGML as a Component of the Digital Library". *Library Hi Tech*, v. 13, n°4, 1995, p. 76.

[114] Jan Corthouts, Richard Philips. "SGML: a librarian's perception". *The Electronic Library*, v. 14, n° 2, 1996, p. 101.

[115] Catherine Lupovici. "L'information secondaire du document primaire: Format MARC ou SGML?". *Bulletin d'Informations de l'Association des Bibliothecaires Francais*, n° 174, 1997, p. 70.

CAPÍTULO II.3

EL DOCUMENTO SGML

II.3.1. PARTES INTEGRANTES E INFRAESTRUCTURA TECNOLÓGICA PARA SU PROCESAMIENTO

SGML interpreta, como ya se comentó con anterioridad, que un documento electrónico puede ser descompuesto en tres elementos claramente diferenciados: estructura, contenido y formato.

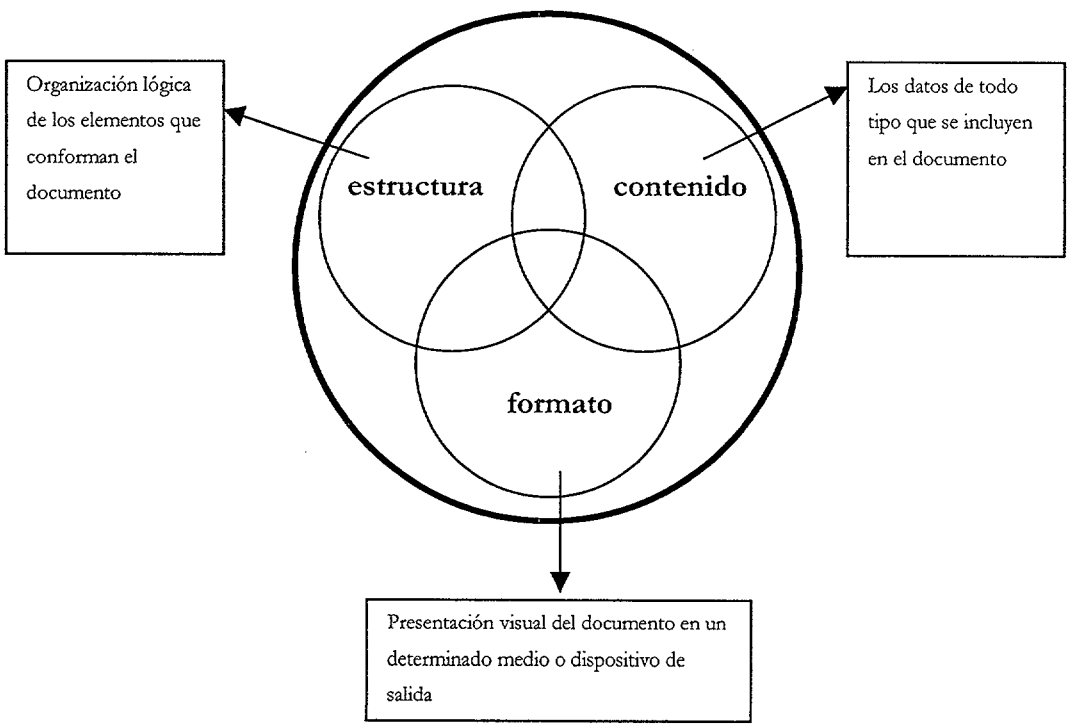
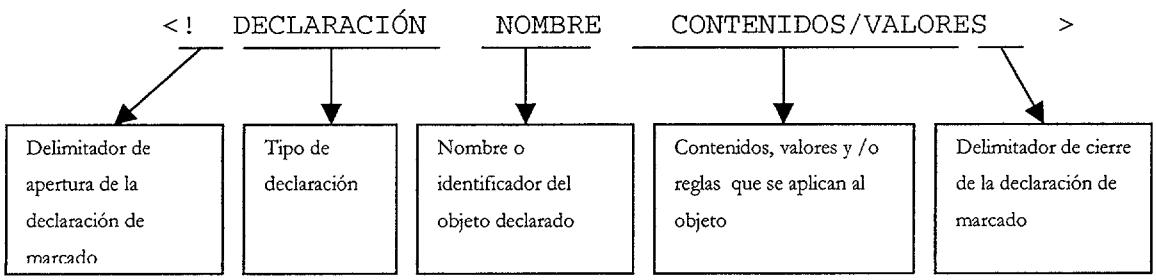


Figura II.4: Descomposición del documento electrónico según el metalenguaje SGML.

Para SGML, de forma abstracta, el término "documento" no hace referencia al concepto tradicional de un conjunto de páginas impresas o a un fichero de ordenador sino, más bien, a una construcción lógica compuesta de una serie de entidades u objetos, donde cada uno

de estos objetos puede contener uno o varios elementos lógicos. A su vez, cada elemento puede contener otros elementos, así como una serie de atributos o propiedades que matizan algunos aspectos de dicho elemento (normalmente, el modo en el que han de ser interpretados y procesados dichos elementos)[116].

Pero, desde un punto de vista más formal y riguroso, cada documento electrónico SGML es una entidad bastante más compleja, compuesta de una serie de declaraciones (o *marcado declarativo*) que informan al sistema informático que ha de procesar el documento sobre ciertos aspectos propios del lenguaje que se está empleando y sobre el tipo de documento al que pertenece el documento en cuestión[117]. Toda declaración en SGML se construye siguiendo unas reglas sintácticas bien sencillas, debiendo contener los componentes básicos que se indican en la siguiente figura:



De forma general, el documento SGML contiene tres partes bien diferenciadas: la declaración SGML, la declaración de tipo de documento y la instancia de documento. Analizaremos brevemente cada una de estas partes fundamentales.

[116] Martin Bryan. *An Introduction to the Standard Generalized Markup Language (SGML)* [documento HTML]. The SGML Centre, 1992. Disponible en http://www.isgmlug.org/sgmlhelp/bryan.htm (consultado el 5 de septiembre de 2000).

1. LA DECLARACIÓN SGML:

La declaración SGML, no obligatoria, incluye instrucciones de procesamiento para los programas informáticos (*parsers*) que analizan este tipo de documentos. En ella se incluye información múltiple sobre diversos aspectos propios del lenguaje de marcado que se está utilizando y de la sintaxis que se está empleando para dicho documento. Como señala E. Peis, esta declaración especifica información básica sobre el dialecto SGML que se está utilizando[118]. Así, además de declararse que el documento se ajusta a la norma SGML de la ISO, se relatan otros aspectos relativos a la utilización de un determinado juego de caracteres, los requerimientos de capacidad máxima que no ha de ser excedida por el documento, la sintaxis concreta para caracteres, funciones, nombres, delimitadores, etc. Todo ello encaminado principalmente al intercambio de datos entre aplicaciones informáticas, la utilización de determinadas opciones dentro del documento como la reducción del marcado, el tipo de enlace hipertextual empleado, el número de tipos de documentos que concurren, cómo han de ser interpretados los identificadores públicos, etc., y, por último, si se añade o se excluye información específica sobre la aplicación.

Debido a la extrema complejidad en la descripción de la declaración SGML, muchos sistemas de procesamiento de documentos SGML permiten su exclusión (aunque su inclusión es siempre recomendable). En el caso de no existir dicha declaración dentro de un documento, el sistema de procesamiento tomará valores por defecto para ciertas variables de procesamiento, como por ejemplo el juego de caracteres, la sintaxis empleada, etc.

Un ejemplo típico de Declaración SGML para un documento SGML básico sería el siguiente:

---

[117] Martin Colby, David S. Jackson. *Using SGML*. Indianapolis, Indiana: Que, 1996, p. 48.

[118] E. Peis, F. de Moya. *Op. cit.*, p. 9.

```
<!SGML  "ISO 8879:1986"
-- This document is a basic SGML document --
-- concrete syntax --
CHARSET
-- 8-bit document character set whose first 128 characters are the same
as the syntax-reference character set. --
BASESET       "ISO 646-1983/CHARSET
              International Reference Version (IRV)//ESC 2/5 4/0"
DESCSET         0     9     UNUSED
                9     2     9
               11     2     UNUSED
               13     1     13
               14    18     UNUSED
               32    95     32
              127     1     UNUSED
BASESET       "ISO Registration Number 109//CHARSET
              ECMA-94 Right Part of Latin Alphabet Nr. 3//ESC 2/9 4/3"
DESCSET       128    32     UNUSED
              160     5     32
              165     1     "SGML User's Group logo"
              166    88     38     -- Includes 5 unused for NONSGML -
              254     1     127    -- Move 127 to unused position as -
              255     1     UNUSED -- 255 is shunned character number -

CAPACITY      PUBLIC "ISO 8879-1986//CAPACITY Reference//EN"
SCOPE         DOCUMENT
SYNTAX        PUBLIC "ISO 8879-1986//SYNTAX Reference//EN"
FEATURES
MINIMIZE      DATATAG NO   OMITTAG  YES  RANK     NO  SHORTTAG YES
LINK          SIMPLE  NO   IMPLICIT NO   EXPLICIT NO
OTHER         CONCUR  NO   SUBDOC   NO   FORMAL   NO
APPINFO       NO>
```

Figura II.5: Declaración SGML de un documento electrónico.

Fuente: Ch. F. Goldfarb, Y. Rubinsky. *The SGML Handbook*, p. 479.

2. LA DECLARACIÓN DE TIPO DE DOCUMENTO:

Precedida por el término DOCTYPE, informa al procesador SGML del nombre y la localización del fichero de texto que define el tipo de documento o DTD (*Document Type Definition*). Aunque el concepto y el proceso de creación de DTDs para documentos SGML se abordarán en detalle en el siguiente punto de este capítulo, sí es necesario hacer una distinción entre la definición y la declaración de tipo de documento, pues a menudo tienden a ser descritos como una misma cosa, cuando en realidad no lo son.

De forma general, se puede decir que la definición del tipo de documento define las reglas de marcado para un tipo concreto de documento, estableciendo los nombres (identificadores genéricos) de los diversos elementos que se integran en éste, los atributos que éstos pueden tomar, así como la relación que se establece entre todos esos elementos[119]; esta definición puede venir completamente descrita dentro del fichero en el que se almacena la instancia de documento (subconjunto interno), completamente descrita en fichero independiente (subconjunto externo) o ser definida tanto interna como externamente.

Un documento SGML se ajusta a un determinado tipo documental a través de la declaración o invocación a la DTD correspondiente; es decir, que el documento ha de ser marcado invariablemente con los elementos y con la jerarquía establecida para éstos, según se ha definido en la DTD. Para declarar el tipo de documento que se está empleando se puede escoger entre una de las diferentes formas establecidas para ello:

- Subconjunto o subparámetro externo de la DTD: la declaración de tipo de documento invoca a la DTD que se encuentra almacenada en un fichero de texto independiente.

  Por ejemplo:

  ```
  <!DOCTYPE  memo  SYSTEM  "C:\DTDS\memo.dtd">
  ```

- Subconjunto o subparámetro interno de la DTD: el desarrollo de la DTD se encuentra contenido dentro de la declaración de tipo de documento.

  Por ejemplo:

  ```
  <!DOCTYPE  memo  [
  <!ELEMENT  memo - - (to, from, subject, p+) >
  <!ELEMENT  (to|from|subject|p)  - 0 (#PCDATA) >
  ]>
  ```

- Definición mixta: la DTD completa está compuesta por la unión de lo definido en el subconjunto externo y lo establecido en el subconjunto interno, permitiendo de este modo tener unas reglas generales de marcado para todos

[119] J. M. Smith. *Op. cit.*, p. 28.

los documentos de un determinado tipo a través del subconjunto externo y complementar éstas en el subconjunto interno con algunas particularidades para un documento determinado.

Por ejemplo:

```
<!DOCTYPE memo SYSTEM "C:\DTDS\memo.dtd"[
<!ELEMENT p - 0 (#RCDATA)>
]>
```

– En caso de conflicto entre el subconjunto externo y el subconjunto interno, tal y como es el caso expuesto en este ejemplo, tendrá mayor peso lo definido en la parte interna de dicha declaración.

3. LA INSTANCIA O MODELO DE DOCUMENTO:

La instancia de documento (también denominada *modelo de documento*) es, como decíamos al principio de este punto, la parte del documento SGML que contiene los datos y todas las marcas, establecidas éstas en la declaración SGML y jerarquizadas según lo impuesto por la definición de tipo de documento declarado. Aquí se incluye el texto del documento junto a todas las etiquetas empleadas para marcarlo, tal como se expuso en el apartado de marcado con SGML (delimitadores de marca, elementos, atributos y referencias a entidades). En algunos casos, la instancia de documento puede venir precedida por un subconjunto de la declaración de tipo de documento, para notificar que ella se hará referencia a entidades adicionales que sólo serán aplicadas a esa instancia de documento concreto (por ejemplo, cuando la instancia del documento lleve imágenes propias y exclusivas de dicha instancia).

Un ejemplo sencillo de instancia de documento ajustado al tipo de documento antes expuesto sería el siguiente:

```
<memo>
<to>Paul M. Ellison
<from>John A. Rivers, Department of Personal
<subject> Denial of the renovation of yours labor contract
<p> Regretting it deeply, I communicate you that this
company won't proceed to the renovation of yours labor
contract.
<p>It is, therefore, work of the Department of Personal to
communicate that you comes for our office any day of next
week to sign your labor ceasing and to pick up your
corresponding monetary quietus.
<p>Without further matter, and awaiting your presence in
our office, I greet sincerely.
</memo>
```

Los componentes que se pueden incluir dentro de un **Sistema SGML** dependen en gran medida de las necesidades a cubrir por la organización, pudiendo ir desde unas sencillas herramientas para la edición de documentos electrónicos SGML hasta complejos y potentes sistemas de bases de datos para la gestión integral de la información electrónica circulante en la organización[120]. Sin entrar en un examen exhaustivo, se puede decir que el establecimiento de la **infraestructura tecnológica** necesaria para la instalación de un sistema de trabajo con documentos SGML conlleva un gran número de operaciones, que se pueden agrupar en tres clases distintas[121]:

1. Operaciones de entrada: incluye procesos como los del análisis de los datos, el desarrollo *ex novo*, la creación, la adaptación o la adopción de una o varias DTDs, así como la entrada de los datos dentro del sistema SGML de almacenamiento.

[120] Dennis J. O'Connor. *The SGML Puzzle: The Pieces and How They Fit Together* [documento HTML]. Mulberry Technologies, [sin fecha]. Disponible en http://www.mulberry.com/papers/puzzle.html (consultado el 9 de septiembre de 2000).

[121] Chris Savage. *SGML: Technical Infrastructure Overview* [documento HTML]. IFLANET, July 1998. Disponible en http://www.ifla.org/VI/5/op/udtop10/udtop10.htm (consultado el 23 de octubre de 2000).

2. Operaciones de procesamiento: son aquellas tareas orientadas al almacenamiento y gestión de los datos SGML, controlando el crecimiento de dichos datos así como los cambios a los que se ven sometidos en el tiempo dentro del proceso de gestión global de la documentación.

3. Operaciones de salida: desarrollo de formatos u hojas de estilo para la presentación de los datos del documento en determinados medios de salida, principalmente pantalla del ordenador o papel impreso.

Esta instalación del sistema SGML requerirá, la más de las veces, una compleja infraestructura tecnológica. Los componentes de *hardware* y *software* que habitualmente se emplean pueden ser agrupados en cinco categorías distintas[122]:

1. Herramientas para el diseño y validación de los datos: etapa inicial de análisis de los datos para determinar y establecer la o las DTDs que mejor se adaptan a la tipología documental de la organización, así como posibles procesos de conversión de los datos ya existentes al formato SGML. Las herramientas informáticas que entran en juego para el perfecto desarrollo de las DTDs son múltiples y variadas: editores, validadores, visualizadores, etc.

2. Herramientas de creación de documentos: una vez que la DTD es elaborada los datos de los documentos deben ser codificados con etiquetas SGML, siguiendo las reglas sintácticas definidas en la DTD. Entre las herramientas informáticas que habitualmente se emplean para esta labor nos encontramos con editores SGML y otro tipo de herramientas para la edición de ficheros multimedia que se puedan integrar dentro de los documentos SGML.

3. Herramientas de conversión de datos en otros formatos: muchas organizaciones almacenan documentos electrónicos en formatos de fichero diferentes,

[122] *Ibid,* http://www.ifla.org/VI/5/op/udtop10/udtop10.htm

normalmente generados por diversos procesadores de texto, por lo que es necesario convertir estos datos al formato propio de SGML. Existen en el mercado herramientas informáticas capaces de realizar estas conversiones, no sin cierta dificultad.

4. Herramientas para el almacenamiento y gestión de los datos: el principal cometido de todo sistema informático para la gestión de documentos SGML está en el almacenamiento, búsqueda, recuperación y mantenimiento de los datos contenidos en esos documentos. El panorama informático es amplio y variado aquí dado que se puede optar por sencillos sistemas de gestión de ficheros hasta caros y complejos sistemas de gestión de bases de datos.

5. Herramientas para la presentación de los datos en múltiples formatos: normalmente los datos SGML contenidos en los sistemas de almacenamiento y gestión, o los propios documentos SGML como tales, necesitan de la definición de estilos para su presentación en diversos medios de salida. Existen en el mercado herramientas capaces de trabajar con lenguajes específicos de estilo para documentos SGML que simplifican notablemente la labor.

Hablaremos, por último, brevemente en este apartado de la noción de analizador de documentos SGML (*SGML parser*) dado que este concepto ya ha sido introducido con anterioridad pero no ha sido explicado suficientemente. La representación textual de los documentos electrónicos marcados es apta para su consulta por parte de los humanos pero, evidentemente, no resulta adecuada para el tratamiento automatizado de dicho texto por parte de los programas informáticos. Se hace, por tanto, imprescindible realizar algún tipo de transformación que verifique la exactitud del contenido incluido en el documento electrónico respecto a las normas a las que debería ajustarse.

Así, un **parser SGML** no es otra cosa, en principio, que un *software* que permite procesar documentos SGML y analizar si éstos se ajustan a lo establecido por la norma ISO 8879. De este modo, el programa chequea el documento SGML para comprobar si es

consistente, si está libre de errores y, en definitiva, si se ha construido de forma correcta[123]. Los *parsers* SGML se pueden presentar de dos formas: como versión limitada para el análisis exclusivo del marcado empleado en el documento SGML o, en la versión más completa y habitual, que analiza tanto el marcado como la conformidad de éste a lo dispuesto por la DTD a la cual referencia. Este último tipo de *parser* se le denomina *parser* validador (*validating parser*). En este último caso, la revisión que realizará dicho programa informático dentro del documento SGML será doble:

- Si la DTD del documento SGML se ajusta a las reglas sintácticas y de construcción establecidas por la norma ISO 8879.
- Si la instancia de documento por tanto, se ajusta a lo detallado en la DTD a la cual está sujeta.

Aunque este programa analizador existe como tal *software* de forma independiente, lo más habitual es que se integre dentro de las herramientas de edición y gestión de este tipo de documento electrónico, ampliando notablemente las posibilidades de trabajo con los mismos (por ejemplo, puede mostrar la estructura jerárquica de los elementos que conforman la DTD, insertar de forma automática aquellas etiquetas cuya presencia es obligatoria, etc.)[124].

En cualquier caso, la norma ISO 8879 contempla en su anexo F, dedicado a la descripción de un posible modelo para un sistema informático SGML, la funcionalidad que debería tener todo analizador o *parser* SGML[125].

---

[123] Eric van Herwijnen. *Practical SGML*. [2ª ed.]. Boston [etc.]: Kluwer Academic Publishers, 1994, p.28.

[124] Steve Pepper, de la empresa noruega STEP Infotek, mantiene un excelente directorio en la Web sobre herramientas de software para trabajar con SGML, con el título de *The Whirlwind Guide to SGML & XML Tools and Vendors*. Para una mayor información sobre este tipo de herramientas informáticas, véase en http://www.infotek.no/sgmltool/

[125] Véase la norma ISO 8879, Anexo F, pp. 154-161.

En definitiva, y a modo de resumen, se puede establecer que SGML proporciona un mecanismo estandarizado para describir tanto los objetos que son contemplados en el documento como las relaciones que se establecen entre sus elementos y los atributos de éstos, e informar al ordenador cómo ha de reconocer y analizar cada una de las partes que componen ese documento[126].

[126] M. Bryan. *Op. cit.*, http://www.isgmlug.org/sgmlhelp/bryan.htm

II.3.2. LA DEFINICIÓN DE TIPO DE DOCUMENTO (DTD)

A lo largo de este capítulo se ha venido insistiendo en el importante peso que trae consigo para los actuales lenguajes de marcado de documentos la descripción de la estructura lógica de los documentos electrónicos frente a la descripción del formato de presentación de los mismos.

Los documentos fuertemente estructurados pueden ser representados como un modelo basado en objetos, en donde los objetos del nivel más alto están compuestos por otros objetos menores. Este hecho permite, según R. Furuta, definir la representación de los documentos electrónicos estructurados a dos niveles distintos: uno genérico y otro específico[127].

El nivel de **estructura lógica genérica** (*generic logical structure*) representaría un modelo para todos aquellos documentos electrónicos que se integran dentro de una determinada clase, familia o tipo documental, y el nivel de **estructura lógica específica** (*specific logical structure*) representaría el modelo concreto de un documento dentro de una clase o tipo documental, plasmado en la instancia de documento de un documento SGML. Esto es, muchos documentos tienen estructuras documentales similares, las cuales pueden agruparse lógicamente y definir tipos documentales concretos (estructuras genéricas). Pero siendo similares dichas estructuras, no son siempre idénticas, y así cada documento dentro de una familia puede tener sus particularidades (estructura específica) que lo diferencian del resto de documentos "hermanos"[128] (por ejemplo, una memoria de investigación puede contener gráficos y otra no). El nivel de estructura lógica genérica debe contemplar todas las

---

[127] Richard Furuta, Vincent Quint, Jacques André. Interactively editing structured documents [documento PDF]. *Electronic Publishing*, v. 1, nº 1, April 1988, p. 22. Disponible en http://cajun.cs.nott.ac.uk/compsci/epo/papers/volume1/issue1/eprxf011.pdf (consultado el 19 de septiembre de 2000).

[128] Peter Fankhauser, Yi Xu. "MarkItUp! An incremental approach to document structure recognition" [documento PDF]. *Electronic Publishing*, v. 6, nº 4, December 1993, p. 450. Disponible en

particularidades que se puedan presentar dentro de cada una de las estructuras lógicas específicas de cada documento que se integre en esa familia.

Tradicionalmente, han sido los mundos profesionales de la Diplomática contemporánea y la Archivística quienes más y mejor han trabajado con el concepto de "tipo documental", dado que para el correcto procesamiento de los expedientes generados por las diversas organizaciones y administraciones ha sido necesario desarrollar una tipología clara de los documentos producidos. Así, remitiéndonos a las palabras de M. P. Martín Pozuelo, se puede definir el concepto de **Tipo Documental** como "la expresión tipificada de documentos con unos mismos caracteres externos, de actuaciones únicas o secuenciales, normalmente reguladas por una norma de procedimiento, derivadas del ejercicio de una función y realizadas por un determinado órgano, unidad o persona con competencia para ello"[129]. Esta misma autora hace hincapié, igualmente, en la idea de que la producción de cada documento responde a una necesidad concreta de la organización para el desarrollo de su labor cotidiana, siendo éstas múltiples y variadas por lo que serán, asimismo, variados los tipos documentales establecidos.

Trasladando esta idea a un terreno más pragmático, es posible comprender fácilmente que los documentos electrónicos que se manejan en muchas instituciones siguen un determinado patrón a la hora de estructurar sus contenidos (por ejemplo, un manual técnico tiene una estructura organizativa de contenidos distinta a una memoria de investigación), por lo que resulta relativamente sencillo agrupar éstos según tipos de documentos. Aunque GML ya introducía este concepto, es con SGML donde alcanza su verdadero significado y potencial. Un tipo de documento es definido formalmente por los elementos lógicos que lo constituyen (señalados con identificadores genéricos) y por las relaciones estructurales y jerárquicas que se establecen entre dichos elementos. Esto es, si el marcado descriptivo de documentos electrónicos, tal como decíamos en apartados

http://cajun.cs.nott.ac.uk/compsci/epc/papers/volume6/issue4/ep6x4pxf.pdf (consultado el 19 de septiembre de 2000).

[129] Mª Paz Martín Pozuelo. *La construcción teórica en Archivística: el principio de procedencia.* Madrid: Universidad Carlos III, Boletín Oficial del Estado, 1996, p. 104.

anteriores, permite dilucidar la estructura lógica del documento que se está marcando, se está indicando asimismo qué elementos se suceden y en qué orden. Por tanto, es posible identificar estructuras lógicas comunes según tipos de documentos diversos y establecer una serie de reglas que definan la estructura permitida para todos los documentos electrónicos que pertenezcan a un tipo de documento determinado.

SGML, como metalenguaje que es, define una sintaxis abstracta para un modelo de lenguaje de marcado generalizado pero, de igual modo, proporciona un mecanismo normalizado para la generación de una familia de lenguajes de marcado descriptivo que pueden ser utilizados para describir la estructura de un documento[130]. Por tanto, junto a la sintaxis abstracta de SGML existe en el propio estándar internacional un mecanismo para especificar de modo normalizado la Definición de Tipo de Documento (DTD).

La DTD definirá la estructura lógica genérica que han de tener los documentos electrónicos que se ajusten a ella (en la instancia o modelo de documento), detallando para ello una serie de propiedades que han de contemplar los elementos que se integran dentro del marcado del documento electrónico, como son los siguientes[131]:

- Los nombres de los elementos que están permitidos.
- La frecuencia de aparición de cada uno de los elementos.
- El orden de aparición de los elementos.
- Las circunstancias en las que las etiquetas iniciales y/o finales pueden ser omitidas.
- El contenido permitido para cada uno de los elementos.
- Los atributos que pueden definir algunas características particulares del elemento.
- Los nombres de todas las entidades que van a ser utilizadas.
- Las particularidades con respecto a convenciones tipográficos especiales (*short references*) que se van a emplear para facilitar el proceso de marcado.

[130] D. Barron. *Op. cit.*, p. 8.

[131] E. v. Herwijnen. *Op. cit.*, pp. 32-33.

- Y, por último, la referencia a la sintaxis concreta para especificar los caracteres específicos del teclado que van a ser utilizados para introducir las marcas o etiquetas en el documento electrónico, así como el modo para redefinir estos caracteres.

Es importante señalar, aunque no nos detendremos mucho en este punto, que las características fundamentales de toda DTD pueden ser formalizadas conceptualmente utilizando un tipo especial de gramática, la denominada *Extended Context Free Grammars* (ECFGs)[132]. El uso de este tipo de gramáticas está muy extendido en el campo de la computación debido a que los *parsers* o analizadores sintácticos se construyen siguiendo una determinada gramática formal. Este tipo de gramáticas fueron definidas a finales de los años 50 por el lingüista Noam Chomsky, cuyos trabajos han tenido tanta importancia y repercusión en el campo de las matemáticas aplicadas a la computación. Uno de los tipos de gramáticas definidas por Chomsky se conoce por el nombre de gramáticas de contexto libre (*context-free grammars*), entre las que se incluyen los modelos gramaticales siguientes: *Backus-Naur Form*, *van Wijngaarden Form* y las ya mencionadas *Extended Context-Free Grammars*[133].

Pero para el usuario final (y para las pretensiones de esta tesis doctoral) que ha de representar un determinado tipo de documento, la DTD se redactará en texto plano o ASCII, utilizando la sintaxis que establece SGML para la identificación de los elementos, atributos y entidades que se integran en el documento[134].

---

[132] Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini. *Representing SGML Documents in Description Logics* [documento PostScript]. Roma: Dipartamento di Informatica e Sistemistica, Università di Roma, 1996. Disponible en http://www.dis.uniroma1.it/pub/degiacomo/dl96sgml.ps.gz (consultado el 19 de septiembre de 2000).

[133] Para una mayor información sobre el uso de gramáticas formales para la creación de analizadores sintácticos en el campo de la computación, véase la obra de Dick Grune, Ceriel Jacobs. *Parsing Techniques: A Practical Guide* [documento PDF]. Amsterdam: Department of Mathematics and Computer Science, Vrije Universiteit, September 1998. Disponible en ftp://ftp.cs.vu.nl/pub/dick/PTAPG/BookBody.pdf (consultado el 19 de septiembre de 2000).

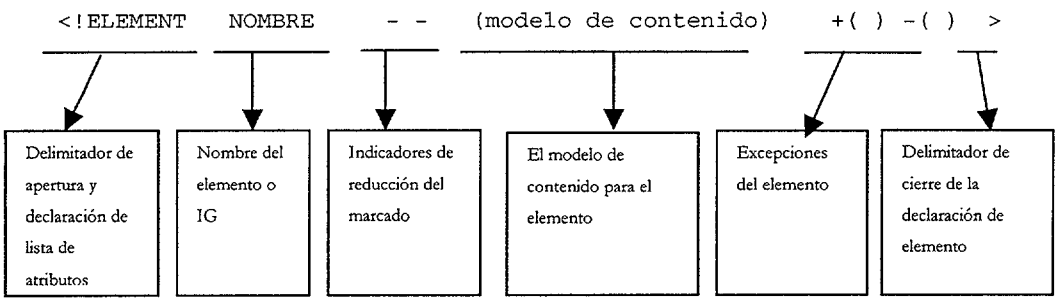
[134] Neil Bradley. "SGML concepts". *Aslib Proceedings*, v. 44, nº 7/8, July/August 1992, p. 273.

A la hora de redactar una DTD, y siguiendo lo impuesto por la norma SGML, se contemplan normalmente tres tipos principales de declaraciones[135]:

1. **DECLARACIÓN DE ELEMENTO** (*element declaration*):

En la cual se definirán los identificadores genéricos o nombres de elementos para cada uno de éstos, indicaciones con respecto a la reducción o minimización de etiquetas iniciales y finales, el modelo de contenido para cada uno de los elementos y las excepciones de inclusión o exclusión de dicho elemento.

La siguiente figura ilustra los componentes que se integran dentro de una declaración de elemento:



El **nombre del elemento** o identificador genérico puede tener cuantos caracteres se desee aunque es recomendable utilizar nombres breves y autoexplicativos. Dicho nombre (*name*) se divide en dos grupos: el primero, denominado carácter de comienzo de nombre (*name start character*), es el primer carácter del nombre y sólo puede ser una letra, mayúscula o minúscula (a-z, A-Z); el segundo, los caracteres del nombre (*name characters*) constituidos por los restantes caracteres, pueden incluir, y combinar, letras mayúsculas o minúsculas, números, el signo de la coma, del punto o el guión (a-z, A-Z, 0-9, , , ., -). Así, ejemplos válidos de nombres de elementos serían: nota, Nota, NOTA, noTa-interna.1, etc.

[135] Ch. F. Goldfarb, Y. Rubinsky. *Op. cit.*, p. 26.

La reducción o minimización de las etiquetas que se insertarán en la instancia del documento SGML permite simplificar el marcado de dicho documento. El signo "**-**" indica que es obligatoria la inclusión de la etiqueta y el signo "**0**" indica que dicha etiqueta puede ser omitida. Así, dependiendo de la posición ocupada por estos signos y por su tipo, se puede establecer la necesidad o no de la inclusión de cada una de las etiquetas, inicial y final, para dicho elemento[136].

El **modelo de contenido** de un elemento define sus subelementos y el orden en que se aceptan, así como las cadenas de caracteres que se pueden suceder en el contenido de ese elemento. En sí, un modelo de contenido es un tipo de grupo compuesto de una serie de elementos conectados entre sí, denominadas *tokens*. Cada grupo de elementos está encerrado entre paréntesis, denominados delimitadores de grupo (de apertura y cierre). Dentro de este grupo de elementos existen otros dos tipos de delimitadores:

- *Conectadores* (, & |): indican la relación que se establece entre los elementos de ese grupo. El conectador "**,**" establece una secuencia o lista de elementos en la que todos se suceden en el orden establecido; el conectador "**&**" indica la aparición de todos los elementos pero en cualquier orden; y el conectador "**|**" establece la unión de dichos elementos, por lo que unos, otros o todos los elementos del grupo podrían aparecer.
- *Indicadores de ocurrencia o aparición* (? + *): establecen cuántas veces pueden aparecer en la instancia del documento SGML cada uno de los elementos contenidos en el grupo. Así, el indicador "**?**" establece que el elemento es opcional, pero si aparece sólo lo puede hacer una vez (0 o una vez); el indicador "**+**" establece que el elemento es obligatorio y que puede aparecer

[136] Algunos autores han criticado las posibilidades que proporciona SGML para la reducción o minimización del marcado de documentos electrónicos debido a los problemas interpretativos que este hecho conlleva, tanto para el usuario como para el programa informático que ha de interpretar una determinada DTD. Para una mayor información sobre este tema, véase el artículo de Sandra A. Mamrak, J. A. Barnes. "Considerations for the preparation of SGML document type definitions" [documento PDF]. *Electronic Publishing*, v. 4, nº 1, March 1991, pp. 27-42. Disponible en http://cajun.cs.nott.ac.uk/compsci/epo/papers/volume4/issue1/ep038.pdf (consultado el 19 de septiembre de 2000).

tantas veces como sea necesario (una o más veces); y el indicador "*" establece que el elemento es opcional, pero de aparecer puede hacerlo tantas veces como sea necesario (0 o más veces).

En algunos casos, el elemento puede contener sólo texto del documento, sin otros subelementos. Para especificar este hecho se reserva (esto es, ningún elemento puede tomar este nombre) el nombre de PCDATA (*Parsed Character Data*) precedido del delimitador "#" , denominado indicador de nombre reservado (*reserved name indicator*). Cuando un elemento declara en su modelo de contenido que contiene tanto otros subelementos junto a texto del documento (#PCDATA), se dice que el contenido de ese elemento es mixto (*Mixed*).

Por último, existen ocasiones en las que partes del documento han de ser procesadas por el *parser* de un modo especial (por ejemplo, una imagen digital, un texto formateado de una determinada forma o un texto que se desea que los signos de marcado permanezcan visibles y no sean, por tanto, procesados). En cada uno de estos casos se reservan una serie de nombres para este proceso, en concreto los siguientes:

- **EMPTY**: puede ser que un elemento no contemple un modelo de contenido; esto es, que el elemento esté vacío; esto es, que en el marcado del elemento no habrá contenido textual encerrado entre la etiqueta inicial y final ni otros subelementos. De hecho, cuando un elemento se declara vacío en la instancia del documento aparece una única etiqueta o *tag*, la etiqueta inicial, en donde el delimitador de cierre vendrá precedido de una barra oblicua (/>) para indicar esta situación especial, tal como se comentó en su momento. Pero el hecho de que un elemento se declare vacío no implica que no contenga información procesable por el programa informático, tan sólo que dicha información será detallada a través de sus atributos. Un ejemplo clásico es el de la inclusión de una imagen digital dentro de un documento SGML: la imagen viene representada por una cadena de bits que definen el tamaño y color de cada uno de los píxeles que la componen. Este conjunto de *bits* tiene un significado distinto a los caracteres de texto y de las marcas, por lo que tendrá que ser analizado e interpretado por el *parser* de modo distinto (de hecho, tal y como

se verá posteriormente, la imagen digital se almacena en una entidad separada cuyo nombre vendrá dado por un atributo especial).

- **CDATA** (*character data*): el contenido existente entre la etiqueta inicial y final de un elemento declarado como CDATA no será analizado e interpretado por la aplicación informática, por lo que si dicho texto incluye los signos propios del marcado SGML (excepto los delimitadores de etiqueta de cierre "</") éstos no se interpretarán y serán mostrados como otros caracteres más.
- **RCDATA** (*replaceable carácter data*): el contenido declarado aquí es similar al de CDATA, excepto que aquí tanto las entidades referenciadas y las referencias a los caracteres sí serán interpretados. Muy utilizado para notaciones especiales como son las ecuaciones matemáticas.
- **ANY**: SGML permite establecer que el contenido de un elemento puede ser cualquier cosa de las ya detalladas, a través del nombre ANY. Se trata de algo, aunque permitido, bastante poco recomendable debido a los problemas que puede generar a la hora de ser interpretado por el *parser*.

Las **excepciones del elemento** (inclusión o exclusión del mismo) permiten al autor establecer que un determinado elemento, que en principio no entra dentro de la lógica formal de la estructura jerárquica establecida para los subelementos de un determinado elemento, pueda ser incluido o excluido en ciertas ocasiones. El delimitador "**+**" permite establecer excepciones de inclusión de elementos dentro del modelo de contenido de un elemento dado y el delimitador **–** excluye al elemento del modelo de contenido del elemento en el cual se introduce.

Por ejemplo:

```
<!ELEMENT  memo - - (to, from, subject, p+) +(foot)>
```

El elemento que configura el nodo superior del árbol de elementos declarados en una DTD, y que por tanto define de forma global el tipo de documento que se está describiendo, se denomina **elemento de documento** (*document element*).

2. DECLARACIÓN DE LISTA DE ATRIBUTOS (*attribute definition list declaration*)[137]:

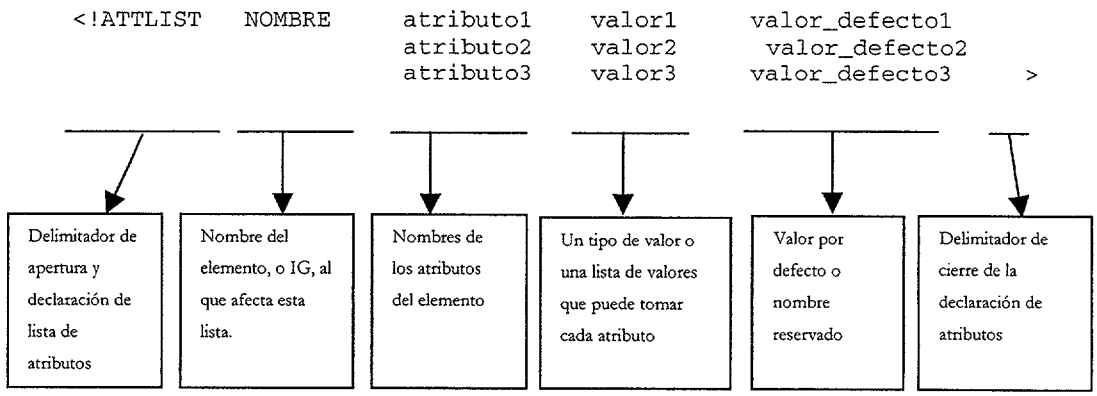
En el entorno de SGML es posible hacer uso de los atributos cuando sea necesario especificar información descriptiva adicional sobre los elementos, normalmente asociando determinados valores a los atributos que pueda llevar un elemento específico[138]. Los atributos que un determinado elemento puede contemplar vienen dados por una declaración formal dentro de la DTD, la cual establecerá el rango de valores que cada atributo puede llevar, así como el valor que se establece por defecto. Si no existe una lista de definición de atributos para un determinado elemento implicará que dicho elemento no contempla el uso de atributos a la hora de marcar un documento electrónico. En resumen, mediante esta declaración se definen los atributos que están permitidos para cada uno de los elementos (en el caso de que los lleve), y su tipo de contenido. También es posible que el contenido que pueda llevar asociado un atributo esté sujeto a unos posibles valores fijos y, por extensión, definir el valor que será tomado por defecto. Ya se comentó en apartados anteriores que esta declaración de atributos tendrá su representación en el marcado del documento electrónico a través de la incorporación del nombre del atributo a la etiqueta inicial del elemento al cual detalla seguida del indicador de valor "=" y del valor del atributo, encerrado entre los delimitadores literales que representan las comillas iniciales y finales.

La siguiente figura ilustra los componentes que se integran dentro de una declaración de atributos dentro de una DTD:

---

[137] El empleo y definición de atributos dentro de los elementos es un tema que ha conllevado ríos de tinta entre los diversos especialistas en la materia dado que no siempre es fácil determinar si un elemento debe llevar o no información asociada que matice ciertos aspectos del texto del documento al cual está marcando. Aunque no existen reglas fijas para determinar este aspecto, suele ser norma común entre los desarrolladores de DTDs la utilización de atributos en las mismas para la normalización del contenido de ciertos elementos (por ejemplo, fechas, nombres propios, etc.) o, también, cuando la descripción de los datos se deba realizar a través de la selección de una serie de valores posibles dentro de una lista (por ejemplo, el nivel de confidencialidad de un documento o la fase de elaboración en la que se encuentra dentro de un proceso de flujo de trabajo). Para una mayor información sobre este tema recomendamos la lectura de la obra de Eric van Herwijnen. *Op. cit.*, pp. 61-62.

[138] M. Colby, D. S. Jackson. *Op. cit.*, p. 182.



El **nombre del atributo** sigue las mismas reglas sintácticas SGML ya vistas para el nombre de los elementos, y de igual forma suele ser un nombre representativo (sustantivos o adjetivos) de su función.

El **valor del atributo** suele ser uno de una lista de posibles valores encerrados entre paréntesis y separados por la barra vertical, de los cuales podrá en su caso señalarse uno de ellos como valor por defecto. Pero también, y en la mayor parte de los casos, es posible declarar este valor según una serie de nombres reservados, indicativos de los diversos contenidos que pueden ser contemplados; éstos son:

- **CDATA**: datos de carácter.
- **ENTITY**: el nombre de una entidad declarada previamente en la DTD.
- **ENTITIES**: una lista de entidades declarada previamente en la DTD.
- **ID**: identificador con valor único de interés para la navegación en el documento (establecimiento de referencias cruzadas dentro del documento).
- **IDREF**: referencia a un identificador.
- **NAME**: un nombre formado sólo por caracteres aceptados para nombres de elementos.
- **NAMES**: una lista de nombres.
- **NMTOKEN**: similar al *name*, establecido por la posible combinación de caracteres alfabéticos, numéricos y los signos del punto, la coma y el guión, pero no siendo obligatorio que comience por un carácter alfabético.

- **NMTOKENS**: una lista de *nmtoken*s.
- **NOTATION**: una notación declarada anteriormente en la DTD.
- **NUMBER**: un nombre formado exclusivamente por caracteres numéricos.
- **NUMBERS**: una lista de números.
- **NUTOKEN**: un nombre *token* establecido por la posible combinación de caracteres alfabéticos, numéricos y los signos del punto, la coma y el guión, pero comenzado siempre por un carácter numérico.
- **NUTOKENS**: una lista de *nutoken*s.

El **valor por defecto** de un atributo puede ser uno de los siguientes[139]:

- Si se han declarado valores diversos para un atributo, uno de ellos puede ser establecido como valor por defecto. En este caso dicho valor por defecto irá entrecomillado.
- Un nombre reservado para indicar una determinada característica del valor del atributo establecido, siendo estos nombres reservados los siguientes (junto con el indicador de nombre reservado):
  - **#FIXED**: el valor establecido para el atributo siempre será el mismo en el documento SGML.
  - **#REQUIRED**: el atributo establecido y su valor definido siempre será obligatorio en el documento SGML.
  - **#CURRENT**: empleado para cambiar el valor establecido por defecto para un atributo dentro del documento SGML. Si un atributo es declarado como *current*, el valor por defecto cambiará automáticamente al valor especificado en último lugar. Esto permite que el valor de un atributo pueda heredar por defecto el valor del elemento del mismo tipo que le ha precedido.

[139] E. v. Herwijnen. *Op. cit.*, p. 92.

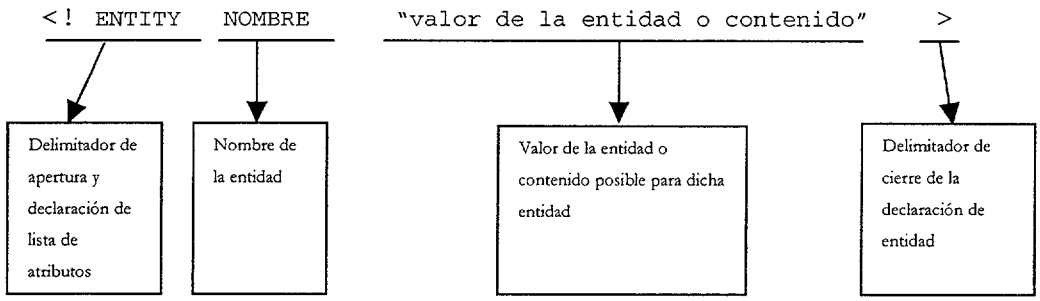
- **#CONREF**: el valor será utilizado para establecer referencias cruzadas entre elementos (especialmente asociado al valor IDREF de un atributo). La aplicación informática de procesamiento puede utilizar este valor del atributo para generar las referencias cruzadas oportunas dentro del documento.
- **#IMPLIED**: el valor del atributo es opcional, por lo que si no es suministrado por el usuario la aplicación informática de procesamiento suministrará uno por sí misma.

3. DECLARACIÓN DE ENTIDAD (*entity declaration*):

Las entidades proporcionan un potente mecanismo para la utilización de notaciones simbólicas dentro de documentos SGML, como ya se comentó en apartados anteriores. El uso de entidades tiene una especial significación y utilidad para los siguientes propósitos[140]:

- Para hacer referencia a caracteres especiales que no están disponibles en los teclados convencionales de los ordenadores (por ejemplo, símbolos matemáticos o caracteres de lenguas no románicas).
- Para la sustitución de cadenas largas de caracteres a través de una referencia breve.
- Para la inclusión o reutilización de fragmentos de documentos electrónicos.
- Para establecer comandos específicos de procesamiento informático a determinados elementos del documento.
- Para señalar la ubicación de ficheros específicos dentro del sistema informático.

De forma general, los componentes que se integran dentro de una declaración de entidad dentro de una DTD pueden ser representados de la siguiente forma:



Existen dos tipos básicos de entidades: las entidades generales y las entidades paramétricas, y ambas pueden ser internas o externas a la DTD.

Las **entidades generales**, referenciadas normalmente[141] en la instancia del documento SGML con los delimitadores inicial del *ampersand* "**&**" y final del punto y coma "**;**", pueden ser internas o externas:

1. El contenido de una **entidad interna** es declarado en la propia DTD del documento SGML, de tal modo que dicho valor declarado será sustituido por el programa informático de procesamiento cuando se encuentre la referencia a dicha entidad. Por ejemplo, la declaración de entidad `<!ENTITY SGML "Standard Generalized Markup Language" >` implica que el nombre de entidad SGML será sustituido por *Standard Generalized Markup Language* cuando la aplicación informática encuentre la referencia a dicha entidad en la forma de &SGML; dentro de la instancia de documento. Normalmente estas entidades son analizadas directamente por el *parser* SGML, comprobando de este modo si su contenido se ajusta o no a lo especificado por la norma.

[140] M. Colby, D. S. Jackson. *Op. cit.*, p. 185.

[141] Y decimos normalmente dado que es posible referenciar o llamar a una entidad general dentro de la instancia del documento a través de un atributo de un elemento. Un atributo cuyo valor declarado en la DTD sea una entidad y dicho valor corresponda en la instancia de documento al nombre asignado a la entidad,

2. Una **entidad externa** puede ser declarada en la DTD. Dicha entidad puede estar almacenada de forma local en nuestro ordenador o ser una entidad normalizada definida por alguna institución pública. Para las entidades externas locales se requiere un identificador de sistema, compuesto por el nombre reservado de SYSTEM seguido de una cadena literal que describe la ubicación y el nombre de los datos del sistema, todo ello entrecomillado. Por ejemplo, la declaración de entidad `<!ENTITY chap SYSTEM "/home/docs/chap.sgml" >` implica que la aplicación informática colocará el contenido del fichero *chap.sgml* (ubicado en el subdirectorio docs del directorio home) en el documento cuando encuentre la referencia a dicha entidad en la forma de *&chap;* Para las entidades externas públicas se utiliza el nombre reservado PUBLIC seguido de una cadena literal que describe de forma normalizada la información necesaria para interpretar dicha entidad. En la mayor parte de los casos es necesario indicar al sistema que las entidades externas que se están declarando contemplan un tipo de datos u otros; esto es, si son datos de carácter o de otro tipo, como datos binarios.

Cuando se declara una entidad general se suele hacer mención del tipo de datos que contiene. Para estos casos se reservan los siguientes nombres CDATA (*character data*) , entidades que sólo contendrán datos de caracteres, por ejemplo `<!ENTITY doc1 SYSTEM "totmemo.sgml" CDATA SGMLdoc >`, SDATA (*special character data*), para entidades de caracteres especiales de datos que dependen del sistema, aunque su utilización es poco frecuente, y NDATA (*non-SGML data*), para entidades cuyo datos sean del tipo binario, como ficheros de imágenes, sonido, vídeo, etc., y que, por tanto, no han de ser directamente analizadas por el *parser* SGML y sí por otra aplicación que el sistema informático tenga designada (en la mayoría de los casos el propio software de visualización de documentos SGML lo posibilita), por ejemplo `<!ENTITY logotipo SYSTEM "logo.gif" NDATA GIF >`.

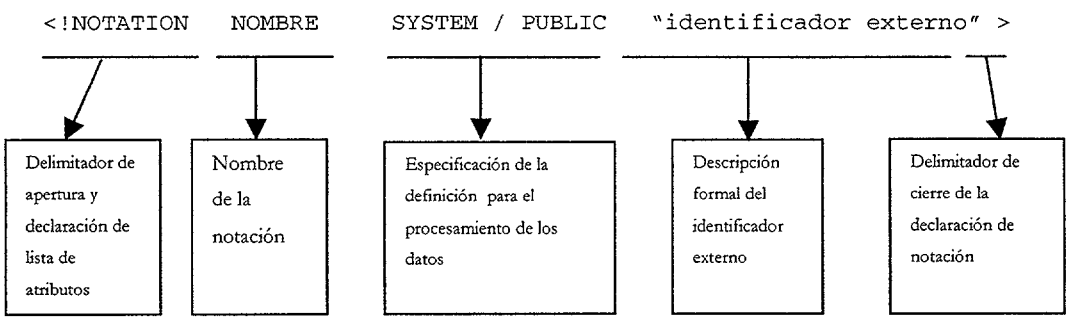
dicha entidad será mostrada cuando se visualice el documento SGML, tal y como se ha explicado anteriormente en el apartado de definición de atributos en la DTD.

Las **entidades paramétricas** son similares a las entidades generales en su definición y propósito (esto es, pueden ser internas o externas, públicas o de sistema), pero sólo pueden ser utilizadas o referenciadas dentro de la propia DTD, nunca en la instancia del documento SGML. La declaración de entidad paramétrica se diferencia de la general en que la primera antepondrá el símbolo del tanto por ciento (%) antes de declarar el nombre de la entidad, en la forma siguiente: `<!ENTITY % nombre "valor de la entidad o contenido" >`. Se trata de un tipo de declaración muy utilizada en las DTDs extensas pues ello permite reducir el texto insertado para la declaración de múltiples partes de la DTD que tienen características comunes entre ellas[142]. Este tipo de entidad también es muy utilizada para hacer establecer entidades públicas que van a ser utilizadas en la DTD, como puede ser, por ejemplo, la referencia a entidades numéricas y gráficas definidas por un organismo de normalización como es la ISO, quedando dicha declaración del siguiente modo `<!ENTITY % ISOnum PUBLIC "ISO 8879-1986//ENTITIES Numeric and Special Graphic//EN" >`. Cualquier referencia a una entidad paramétrica dentro de la DTD vendrá definida por el delimitador inicial del tanto por ciento "%" y por el delimitador final del punto y como ";".

Aunque se han visto y comentado aquí los tres tipos principales de declaraciones que se dan dentro de una DTD hay que decir que en ciertos casos resulta necesario realizar una cuarta declaración, la **DECLARACIÓN DE NOTACIÓN**. El hecho de que SGML fuese creado como lenguaje de marcado de documentos de tipo textual implica que no exista, de forma directa, un soporte para otro tipo de contenido de datos, tales como imágenes digitalizadas, gráficos, fórmulas matemáticas, audio, vídeo, etc.; es decir, para documentos multimedia. Aunque una entidad de datos de carácter o binario puede ser procesada directamente por el sistema, como se ha visto anteriormente, a veces resulta conveniente incluir dentro de la instancia de documentos los datos que

[142] Por ejemplo, una lista de atributos puede ser común a diversos elementos por lo que si dicha lista se define como una entidad paramétrica, cada declaración de elemento puede contener una referencia a esta entidad y no tener así que redactar todo el contenido de dicha lista.

representan a estos objetos especiales (fórmulas, texto formateado, datos binarios, etc.) expresados en su propio lenguaje de producción. De este modo, deberá existir una declaración de notación para cada uno de estos objetos especiales declarando el modo correcto de procesar dichos datos[143]. El identificador de esta notación puede ser, al igual que sucedía con las entidades, SYSTEM o PUBLIC. De forma general, los componentes que se integran dentro de una declaración de notación dentro de una DTD pueden ser representados de la siguiente forma:



Por otro lado, si se ha declarado una notación dentro de la DTD debe existir igualmente dentro de la declaración de atributos del elemento que vaya a hacer uso de dicho contenido de datos no SGML la referencia a dicha utilización, haciendo uso para ello, como se vio anteriormente, del nombre reservado de NOTATION dentro de la declaración del atributo correspondiente. Así, poniendo como ejemplo la inclusión de una fórmula matemática defina a través del TeX (lenguaje especialmente diseñados para esta función), en la DTD se debería reflejar del siguiente modo:

```
<!NOTATION TeX  SYSTEM " ">
<!ELEMENT formula - - CDATA >
<!ATTLIST formula datos NOTATION (TeX) #REQUIRED >
```

El documento SGML quedaría marcado en su instancia de documento del siguiente modo:

[143] E. v. Herwijnen. *Op. cit.*, p. 161.

```
<formula datos="TeX">
$$ { \Gamma (J/\psi \rightarrow \eta_c \gamma) } = { { \alpha
Q_c^2 } \over 24 } \left | A ( J/\psi \rightarrow \eta_c
\gamma) \right|^2 { { m_\psi^3 } \over { m_{ \eta_c}^2} }
\left (1 - { { \eta_c}^2 } \over {m_\psi ^2 } } \right )^3$$
</formula>
```

De este modo, cuando la aplicación SGML llegue a este punto recurrirá al sistema para encontrar el programa capaz de interpretar esta formulación y lo mostrará en un medio de salida (generalmente la pantalla del ordenador) como una representación matemática convencional. De forma parecida se procedería para representar gráficos definidos, por ejemplo, en el lenguaje PostScript, y otro tipo de objetos multimedia. Como es lógico suponer, no es el mejor método para representar ficheros multimedia complejos dado que las notaciones que habría que generar serían realmente extensas. Es por esta razón que para representar objetos multimedia se recurra preferentemente al uso de entidades binarias y dejar que el visualizador de documentos SGML se encargue de presentar en el medio de salida correspondiente dichos objetos.

Por último, dentro de la DTD es posible incluir **comentarios** para realizar aclaraciones o anotaciones sobre el significado y uso de ciertos elementos, atributos y/o entidades, o suministrar cualquier tipo de información de interés para futuros autores de documentos SGML que tengan que hacer uso de esta DTD. Al igual que ocurría con el marcado de la instancia de documento, las marcas empleadas para señalar estos comentarios serán "<!--" al principio y "-->" al final.

Un ejemplo sencillo pero ilustrativo de una DTD SGML para un documento del tipo de notas internas dentro de una institución, junto al documento marcado siguiendo lo especificado en dicha DTD, sería el siguiente:

- Analizando la *estructura lógica* que pueden conformar todas las notas internas de una determinada institución, dicha estructura podría normalizarse del siguiente modo:

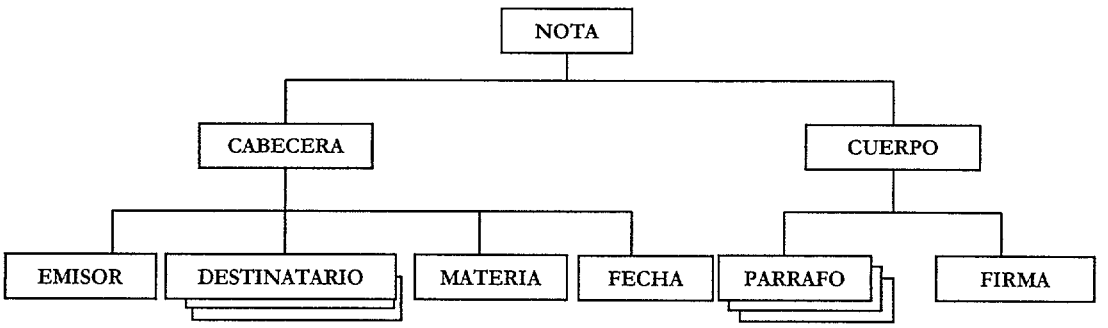


Figura II.6: Posible estructura lógica de una nota interior.

- La *Definición de Tipo de Documento* ajustada a esta estructura lógica que conforman las notas internas, sería la siguiente:

```
<!ELEMENT nota - - (cabecera, cuerpo) >
<!ATTLIST nota  estado  (final|borrador)  "final"
                seguridad  CDATA  #REQUIRED >
<!ELEMENT cabecera - - (emisor, destinatario+, materia,
fecha) >
<!ELEMENT emisor  - 0  (#PCDATA) >
<!ELEMENT destinatario  - 0  (#PCDATA) >
<!ELEMENT materia - 0  (#PCDATA) >
<!ELEMENT fecha - 0 EMPTY >
<!ATTLIST fecha entrada    NUMBER  "01012000" >
<!ELEMENT cuerpo - -  (parrafo+, firma) >
<!ELEMENT parrafo - 0  (#PCDATA) >
<!ELEMENT firma - 0 (#PCDATA) >
```

- Por último, cualquier nota interna que se produjese dentro de la institución debe obligatoriamente ajustarse a lo definido para su tipo de documento en la DTD. Por ejemplo, el *documento electrónico marcado* de forma básica quedaría del siguiente modo:

```
<nota estado=final seguridad=baja>
<cabecera>
<emisor> Marcos R. Basmayer
<destinatario>Bonifacio Martín Galán
<materia>Renovación de contrato laboral
<fecha entrada=01102000 />
</cabecera>
<cuerpo>
<parrafo>Nos es grato comunicarle que le ha sido renovado
por un año más el contrato laboral que mantiene con esta
empresa.
<parrafo>Es por este motivo por el que le ruego que en la
mayor brevedad de tiempo se pase por nuestras dependencias a
fin de formalizar dicha renovación.
<parrafo>Sin otro particular, y a la espera de su pronta
visita, se despide atentamente:
<firma>Marcos R. Basmayer, Director del Departamento de
Personal
</cuerpo>
</nota>
```

En resumen, SGML constituye un metalenguaje que permite definir lenguajes específicos para marcar documentos electrónicos de todo tipo. Cada uno de estos tipos documentales viene definido por una DTD particular, lo que constituye un lenguaje específico de marcado de documentos electrónicos válido y exclusivo de dicho tipo de documento; cada tipo de documento tiene una DTD distinta. Por tanto, se contemplan en este modelo tres bloques fundamentales: SGML como metalenguaje para definir lenguajes específicos de marcado, el lenguaje específico definido en una DTD y el documento electrónico final marcado según lo dispuesto en su correspondiente DTD.

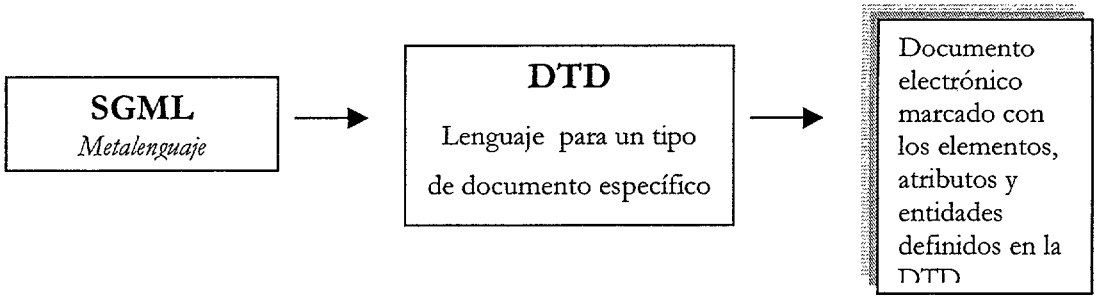


Figura II.7: Bloques fundamentales en el modelo del metalenguaje SGML.

II.3.3. ANÁLISIS Y RECONOCIMIENTO DEL DOCUMENTO PARA LA DEFINICIÓN DE TIPOS DOCUMENTALES

La importancia de la definición de tipos documentales dentro de toda institución resulta un factor clave para una correcta gestión documental, tal y como se ha venido argumentado en apartados anteriores. Ésta viene impuesta por las propiedades que tienen los documentos para transmitir conocimiento pues, de hecho, son el medio tradicional y más idóneo para dicha transmisión.

En los momentos actuales, en que el volumen de documentos electrónicos generados y almacenados por cualquier institución crece día a día de forma notable, se hace imprescindible afrontar los problemas que se derivan de la aplicación de los sistemas informatizados para la producción y gestión de dichos documentos: gran diversidad de formatos de documentos electrónicos, diversidad de plataformas físicas y lógicas, continuas y rápidas transformaciones tecnológicas, etc. Es aquí donde entra en juego el concepto de la *Electronic Document Managment* (EDM) referido al uso de las tecnologías de la información con el propósito de gestionar adecuadamente la producción documental de una organización[144]. Este concepto tan amplio incluye un gran número de tecnologías diversas que toman como punto de referencia a los documentos, tanto en soporte tradicional como en formato electrónico. Estas tecnologías proporcionan, según R. H. Sprague[145], una doble funcionalidad: funciones para el procesamiento de los documentos (como son la captura y creación de documentos, el almacenamiento y organización, diseño de arquitectura para documentos complejos, el almacenamiento distribuido, técnicas de hipertexto, búsqueda y recuperación documental, procesos de transmisión y encaminamiento de los documentos –técnicas de *workflow*-, y presentación de los mismos) y funciones para la gestión de dichos

[144] Ralph H. Sprague. "Electronic Document Management: Challenges and Opportunities for Information Systems Managers". *MIS Quarterly*, v. 19, nº 1, 1995, p. 31.

documentos (informes sobre el estado del documento, controles de acceso, controles de versión, controles para la conservación y expurgo, y controles para la prevención de daños).

Como ya se ha expuesto con anterioridad en este capítulo, el surgimiento de estándares documentales, como fue el caso de SGML, trataba de paliar los problemas derivados de la multitud de formatos diversos de documentos electrónicos dentro de las organizaciones, así como de lograr el modo más eficiente de representación de la información dentro de los documentos que dichas instituciones creaban y gestionaban. Para hacer posible esta tarea en un entorno de producción documental SGML (y, por extensión, XML) es absolutamente necesario, entre otras muchas cosas, entender el papel que juegan los documentos en los procesos de trabajo y, de forma especial, identificar y definir las clases o tipologías documentales que entran en juego en dichos procesos de trabajo para su normalización[146]. Así, surge una etapa de vital importancia en la puesta en marcha de un sistema EDM: el análisis documental.

El análisis documental, desde la óptica de los sistemas SGML, consiste en la revisión y estudio comprensivo de los documentos producidos por una organización, o una muestra de ellos, a fin de deducir tanto el contenido como la estructura de los mismos[147]. En esta fase de análisis documental, como sostiene por A. Salminen, los documentos actuales de la organización, así como los procesos para la gestión de los mismos, son examinados y descritos, lo que producirá diversas definiciones de estructuras lógicas documentales, las cuales, como ya se ha expuesto anteriormente, identificarán los ítems informativos esenciales y las relaciones que se establecen entre éstos dentro de los documentos[148]. De

---

[145] *Ibid.*, pp. 37-39.

[146] A. Salminen, V. Lyytikäinen, P. Tiitinen. "Putting documents into their context in document analysis". *Information Processing and Management*, v. 36, nº 4, 2000, p. 624.

[147] Simon Heathfield. *Standard Generalized Markup Language: Frequently Asked Questions* [documento HTML]. Disponible en http://www.jorvic.demon.co.uk/sgmlfaq.htm (consultado el 7 de septiembre de 1999).

[148] Airi Salminen, Katri Kauppinen, Merja Lehtovaara. "Towards a Methodology for Document Analisis". *Journal of the American Society for Information Science*, v. 48, nº 7, 1997, p. 644.

este modo, si SGML se interesa especialmente por la estructura de los documentos, requerirá, por tanto, un conocimiento profundo de las estructuras físicas y lógicas inherentes de los documentos que se manejan en la organización y, de igual forma, cómo estas estructuras se agrupan o asocian en estructuras lógicas genéricas capaces de describir estructuras lógicas específicas de cada uno de dichos documentos.

Como señalan D. Dori y sus colaboradores, es habitual que la información sea incluida en los documentos a través de cadenas de caracteres, y que estos caracteres vayan conformando objetos cada vez más grandes como palabras, frases, párrafos, etc., y así sucesivamente hasta llegar al nivel completo formado por el documento[149]. De este modo, la estructura de un documento resultaría de la división y subdivisión del contenido del mismo en pequeñas partes, comúnmente denominadas objetos. Cuando un objeto no puede subdividirse en otros objetos más pequeños, a éste se le denomina objeto básico (*basic object*). El resto de objetos toman el nombre de objetos compuestos (*composite object*)[150]. Dependiendo de la naturaleza de la descripción de estos objetos, la estructura podrá ser expresada de una forma u otra: como estructura física o geométrica cuando atiende a las características geométricas que presenta la representación visual del documento o, por el contrario, como estructura lógica cuando se atiende a las propiedades semánticas del mismo. Ambas estructuras estarán, como es de comprender, fuertemente relacionadas, dado que el mismo contenido de un documento puede ser representado con respecto a ambas estructuras.

Por último, volver a recordar que todos los documentos que produce y gestiona cualquier organización pueden ser clasificados según tipos distintos, dado que sus estructuras físicas/lógicas varían de un tipo a otro. Así, cuando se abordan intentos de descripción de relaciones estructurales entre los componentes de un documento, es necesario establecer

---

[149] Dov Dori, David Doermann, Christian Shin, Robert Haralick, Ihsin Phillips, Mitchell Buchman, David Ross. *The Representation of Document Structure: a Generic Object-Process Analysis* [documento PostScript]. Baltimore: University of Maryland, [1996?], p. 11. Disponible en http://lamp.cfar.umd.edu/Media/Publications/Papers/ddori95a/ddoria.ps (consultado el 23 de octubre de 2000).

un mayor nivel de abstracción para, de este modo, dilucidar estructuras lógicas genéricas según el tipo o clase documental al que se adscriban los documentos.

Antes de abordar un proyecto de definición y desarrollo de DTDs en cualquier organización es necesario valorar detenidamente la conveniencia o no de utilizar alguna DTD ya desarrollada por instituciones de prestigio. A estas DTDs se las suele denominar *Industry Standard DTD* (IS DTD), dado que su consenso y creación ha sido fruto de un importante esfuerzo de múltiples personas y, en muchos casos, procedentes de distintas organizaciones de un mismo campo[151]. Su posible utilización dependerá de factores tales como la adecuación de alguna de estas DTDs a las peculiaridades de la tipología documental propia de la organización, la capacidad profesional y la disponibilidad de tiempo de sus miembros para desarrollar una DTD *ad hoc*, los costes económicos asociados a esta tarea si va a ser desarrollada por miembros de la organización o, por el contrario, delegada en una empresa externa de servicios informáticos o documentales, etc.

El gran problema de las DTDs estandarizadas es su generalidad dado que son concebidas para ser empleadas teniendo en cuenta todos las variaciones de elementos que se pueden dar dentro de los documentos de un determinado tipo y que, además, se producen en las diversas organizaciones que han participado en su elaboración. En cualquier caso, como señala S. Heathfield, la elección inadecuada de una DTD estándar puede acarrear los siguientes problemas[152]:

- Obligar a los autores de los documentos a emplear estructuras inapropiadas a las cuales están poco familiarizados.

---

[150] Y.Y. Tang [*et al.*]. *Op. cit.*, p. 6.

[151] Bradley Neil. *The Role of Industry Standard DTDs* [documento HTML]. Barcelona: SGML Europe 97 (1997. Barcelona). GCA, 1997. Disponible en http://www.infoloom.com/gcaconfs/WEB/barcelona97/bradle11.HTM (consultado el 5 de octubre de 2000).

[152] S. Heathfield. *Op. cit.*, http://www.jorvic.demon.co.uk/sgmlfaq.htm

- Forzar los datos de los documentos de la organización dentro de elementos poco apropiados, limitando la capacidad en la indexación y posterior búsqueda y recuperación de la información contenida.
- Proporcionar una amplia elección de elementos posibles lo que podrían generar una alta inconsistencia en el marcado del documento.
- Dificultar la conversión a otros formatos.
- Dificultar la asociación de estilos de salida o presentación de los documentos.
- Obligar a las aplicaciones informáticas a realizar un esfuerzo extra para procesar los datos del documento.
- Aumentar los costes y tiempos para la organización si al final se ve que el procedimiento ha sido inadecuado y es, por tanto, necesario volver a remarcar los documentos de acuerdo a otra DTD distinta.

Por todo ello, suele ser bastante habitual en muchas organizaciones a la hora de desarrollar un proyecto documental basado en SGML tomar como base de partida alguna de las DTDs realizadas por la industria para su estudio y análisis, de lo cual se extraerán a buen seguro importantes observaciones. A partir de ahí, y con la experiencia adquirida, se procederá a tomar todo aquello que pueda resultar de utilidad (división de los grandes bloques, nombres para los identificadores genéricos, modelos de contenido para ciertos elementos, etc.) y se adaptarán a las necesidades particulares de la documentación propia de la organización.

En el caso que la organización decida abordar directamente la construcción de sus propias DTDs, el camino no será mucho menos complejo dado que no existen métodos predeterminados de trabajo para la correcta definición de estructuras lógicas de tipos documentales dentro de las organizaciones, por lo que dependiendo del campo o ámbito científico del que partan los diversos investigadores que han tratado estos temas los planteamientos serán unos u otros. Lo que sí es unánime en todos ellos es la afirmación de que cualquier proyecto de implementación de un sistema documental basado en SGML

debe de partir inicial e inexorablemente de la definición del entorno de trabajo en el que se producen y utilizan los documentos[153].

Uno de los campos científicos que más y mejor ha tratado el tema del análisis documental para dilucidar y normalizar estructuras lógicas de tipos documentales diversos dentro de las organizaciones es el de la aplicación de las técnicas de digitalización de documentos y reconocimiento óptico de caracteres (OCR). Y es lógico que así sea dado que muchos de los grandes proyectos de gestión electrónica documental que se han venido abordando en estas últimas décadas debían comenzar invariablemente por la conversión de la vasta colección de documentos en soporte papel existente a un formato electrónico legible por el ordenador. De este modo, es posible identificar automáticamente grupos de palabras y segmentos de línea que forman las unidades principales de la estructura tipográfica del documento[154]. Además, en esa fase de digitalización y reconocimiento de caracteres es posible, además, intentar dilucidar los elementos que definen a las estructuras lógicas que se derivan de las imágenes producidas por el escáner. De este modo, según D. Niyogi, a través de estas técnicas es posible determinar las relaciones que se establecen entre la representación física de un documento (consistente en la estructura geométrica y relaciones espaciales que se establecen entre los bloques de un documento impreso) y su correspondiente representación lógica (consistente en la agrupación lógica de dichos bloques dentro de unidades contenedoras)[155].

---

[153] Colby y Jackson establecen magistralmente en el capítulo 6 de su obra los principios que se han de seguir para la definición del entorno, los cuales se pueden enumerar en los siguientes puntos: definir quiénes hacen uso de los documentos, de qué modo y para qué; definir qué estándares y políticas van a ser seguidas; identificar todos los usuarios que generan documentos y las herramientas que utilizan para dicha labor; recoger e inspeccionar todos los tipos documentales que se producen, así como asignarles nombres definitorios; anticiparse a la evolución tecnológica que han de sufrir dichos documentos; y corregir y redefinir este entorno documental con los errores que se hayan detectado. Para una mayor información sobre el tema véase Martín Colby, David S. Jackson. *Op. cit.*, pp. 109-127.

[154] David Slocombe, Jyoti Ambekar. "Document Structure Identification: a New Paradigm" [documento HTML]. En: *SGML/XML Europe 98 Conference* (1998. Paris). GCA, May 1998. Disponible en http://www.infoloom.com/gcaconfs/WEB/paris98/slocombe.HTM (consultado el 5 de octubre de 2000).

[155] Debashish Niyogi, Sargur N. Srihari. *The use of Documents structure analysis to retrieve information from documents in digital libraries* [documento PostScript]. Buffalo, NY: Center of Excellence for Document Analysis and

Yuan Y. Tang y otros colaboradores[156], investigadores dentro de este campo de actuación, parten de la idea de que para poder llevar a cabo el procesamiento de los documentos a través de estas técnicas digitales es necesario realizar dos fases bien diferenciadas: un proceso de análisis documental (*document analysis*) y otro de posterior reconocimiento documental (*document understanding*). El proceso de análisis documental iría encaminado a extraer la estructura física o geométrica del documento para, posteriormente, y a través de técnicas de reconocimiento documental, "mapear" ésta para extraer su estructura lógica. La siguiente figura ilustra dicho proceso:

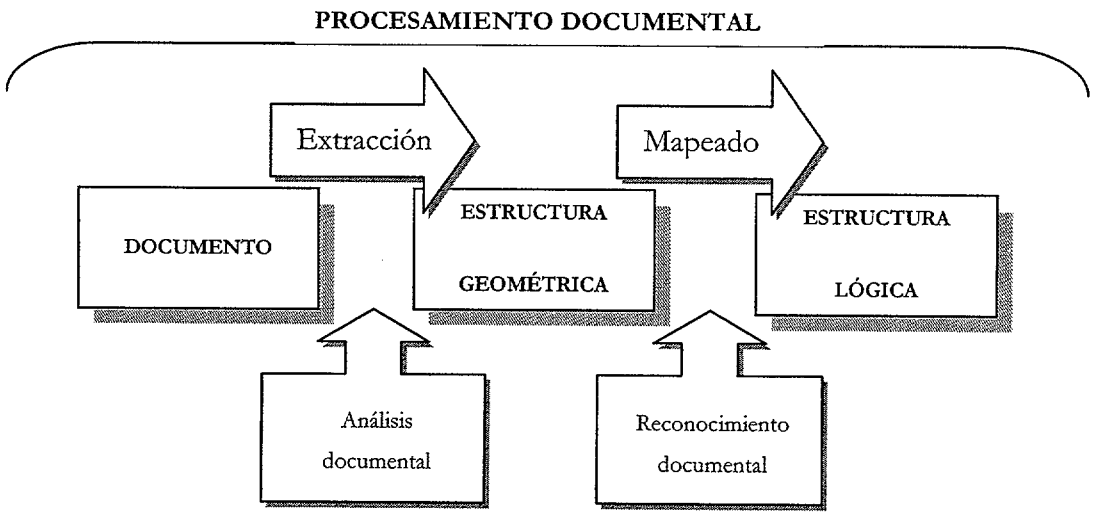


Figura II.8: Fases del procesamiento del documento mediante técnicas de digitalización.

Recognition, State University of New York, 1997, p. 2. Disponible en http://www.cedar.buffalo.edu/~niyogi/papers/SPIE97/spie97.ps (consultado el 23 de octubre de 2000).

[156] Yuan Y. Tang and M. Cheriet, Jiming Liu, J. N. Said, Ching Y. *Document Analysis and Recognition by Computers* [documento PostScript]. Hong Kong: Department of Computing Studies, Hong Kong Baptist Univesity, [1993?], p. 3. Disponible en ftp://robotics.comp.hkbu.edu.hk/pub/doc/handbk.ps.gz (consultado el 9 de octubre de 2000).

El **Análisis Documental** (*document analysis*) es, por tanto, definido en este contexto como la extracción de la estructura física o geométrica de un documento. Este análisis conlleva la disección de la imagen del documento en diversos bloques los cuales definirán cada uno de los objetos que lo componen, tales como bloques de texto, cabeceras, imágenes, etc. La estructura resultante de este proceso de análisis documental puede ser representada como un árbol geométrico (y, normalmente, con el establecimiento jerárquico de dichos bloques) tal y como se muestra en la siguiente figura para el caso, por ejemplo, de un Real Decreto de noviembre de 2000 publicado en el Boletín Oficial del Estado (en este ejemplo se ha omitido deliberadamente el apartado de anexos para no complicar en exceso los gráficos resultantes):

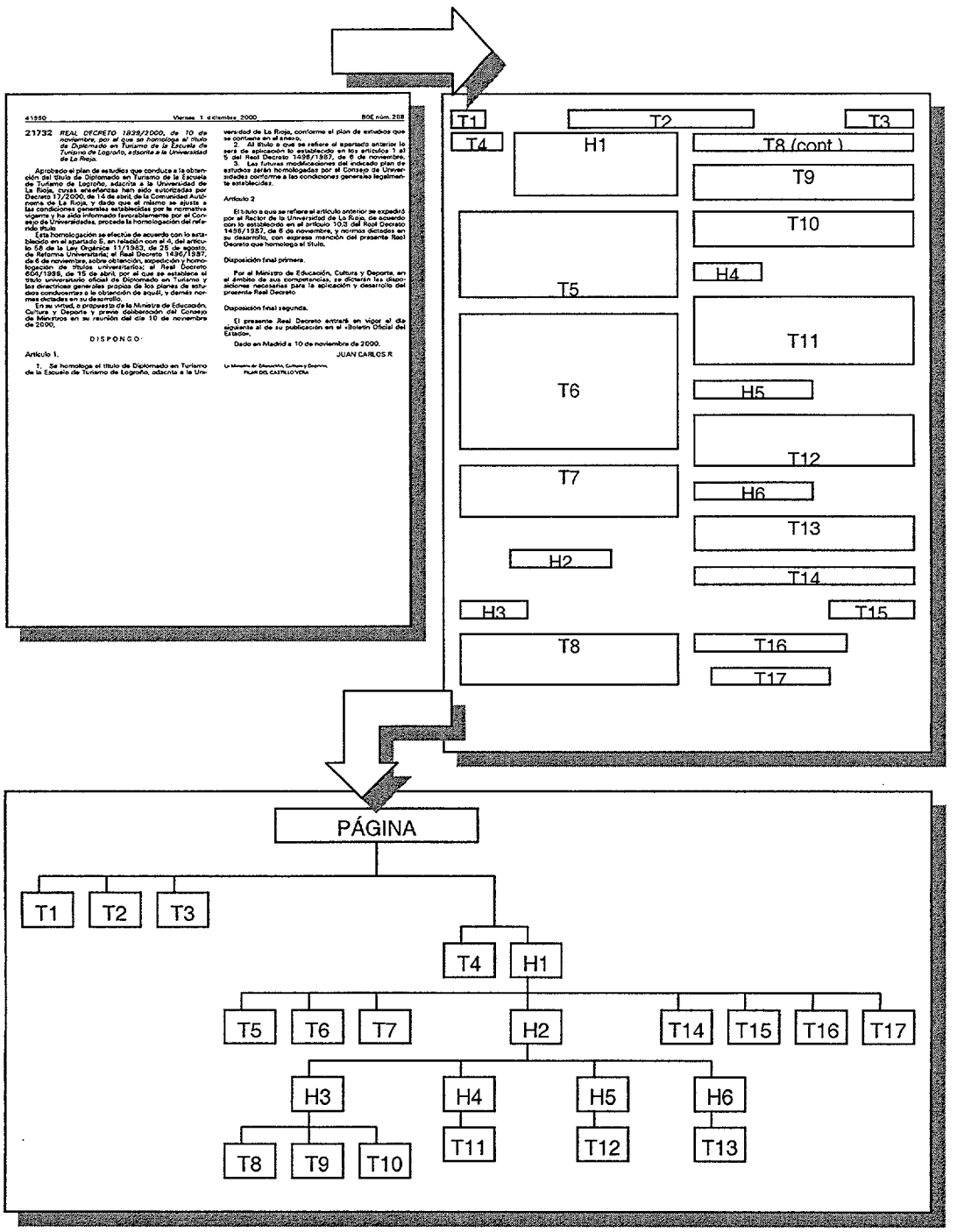


Figura II.9: Descomposición geométrica y árbol jerárquico de elementos de un Real Decreto.

Sin entrar en demasiados detalles, dado que el la digitalización de documentos y el reconocimiento automático de su contenido están fuera de la cobertura temática de esta tesis, sí es necesario comentar que existen diversos mecanismos para la construcción de este árbol geométrico, aunque los más frecuentes suelen englobarse en dos categorías distintas: Métodos jerárquicos y métodos no jerárquicos[157].

1. **Métodos jerárquicos**: Cuando se descompone el documento en diversos bloques las relaciones de supeditación y jerarquía que se establecen entre los mismos resulta ser el factor determinante a la hora de representar el árbol geométrico. Esta construcción se puede realizar, asimismo, siguiendo dos vías distintas: desde los objetos padres hacia los objetos hijos (*top-down approach*) o, por el contrario, desde los hijos hacia los padres (*bottom-up approach*). En el primero de ellos se divide el documento en las principales regiones para posteriormente ir subdividiendo éstas en bloques menores. En el segundo caso se van refinando los datos progresivamente a través de operaciones de agrupación o fusión de objetos según características comunes. Cada uno de estos métodos tiene sus ventajas y sus inconvenientes. Los métodos *top-down* resultan rápidos y muy efectivos para el procesamiento de documentos que tiene un formato específico. En el caso de los métodos *bottom-up* es posible desarrollar diversos algoritmos matemáticos que pueden ser aplicados a un gran número de documentos pero, por el contrario, son métodos más lentos dado que consumen un mayor tiempo de procesamiento.

2. **Métodos no jerárquicos**: tradicionalmente los métodos jerárquicos han sido los más utilizados para determinar la estructura geométrica de los documentos aunque no sin ciertas limitaciones dado que estas técnicas no resultan efectivas para el procesamiento de documentos con una alta complejidad en sus estructuras geométricas, conllevando, asimismo, un gran consumo de tiempo

[157] *Ibid.*, p. 10.

para su realización. Frente a esto surgen los métodos no jerarquizados, los cuales no necesitan procesos de ruptura o de fusión del documento, pero que son capaces de reconocer los diversos objetos a través de algoritmos matemáticos y dividir el documento en bloques en una sola fase. Son especialmente útiles para procesar una gran variedad de tipos documentales incluyendo aquellos que contienen una gran complejidad estructural.

En cuanto al **Reconocimiento Documental**, se trata ésta de la segunda etapa aplicada en estas técnicas de tratamiento de documentos digitalizados y, como ya anunciamos, recoge el árbol geométrico anteriormente producido para "mapearlo" y obtener, de este modo, la estructura lógica del documento, en la cual se establecen las relaciones lógicas entre los objetos contenidos en un documento concreto. Al igual que ocurría en la anterior fase, el método de mapeado para este reconocimiento documental producirá normalmente como resultado un árbol lógico de relaciones de objetos, aunque existen otros métodos para dicha tarea dentro de este campo de actuación que apuntaremos brevemente; a saber[158]:

1. **Reconocimiento documental basado en una transformación arborescente:** este método transforma el árbol estructural geométrico en un árbol estructural lógico. Resulta ser el método más simple y sencillo pues la estructura lógica de un documento es muchas veces definida, tal y como lo expresa K. Summers, como "una jerarquía de segmentos del documento, donde cada uno de ellos corresponde con los componentes semánticos que visualmente se distinguen en el documento"[159]. Durante esta transformación se definen, normalmente de forma automática, una cabecera y un cuerpo, o uno de los dos exclusivamente.

[158] *Ibid.*, p. 16.

[159] Kristen Summers. *Towards a Taxonomy of Logical Document Structures* [documento HTML]. Boston, Massachusetts: DAGS95 – Electronic Publishing and the Information Superhighway, May 30-June 2, 1995. Disponible en http://www.cs.dartmouth.edu/~samr/DAGS95/Papers/summers.html (consultado el 24 de octubre de 2000).

Cada nodo del árbol geométrico es transferido al árbol geométrico asignando una serie de etiquetas dependiendo de la composición y de la posible semántica de cada uno de dichos nodos. Así, las etiquetas que se incluyen pueden ser de título, subtítulo, resumen, cabeceras, párrafos, pie de nota, etc. Las reglas o algoritmos matemáticos de carácter generalista que se aplican para esta transformación están basados en la observación sobre la construcción "lógica" de la mayor parte de los documentos que se manejan en las organizaciones (por ejemplo, un título suele contener un conjunto de párrafos como hijos en la estructura lógica, etc.)[160].

2. **Reconocimiento documental basado en el formateado del conocimiento**: el gran problema de estas técnicas se presenta cuando, como a veces suele suceder, a una estructura lógica le puedan corresponder múltiples estructuras geométricas. En este caso se procede a construir reglas específicas basadas en el formato propio de cada uno de los posibles tipos documentales que se manejan en la organización para, de este modo, obtener una estructura lógica propia de cada uno de esos tipos documentales. Aunque resulta un método algo más complejo si existen numerosos tipos documentales en la organización, los resultados que se pueden obtener son mucho más precisos.

3. **Reconocimiento documental basado en la descripción del lenguaje**: se trata de uno de los métodos más efectivos para describir las estructuras de los documentos. Para ello hace uso de reglas de conocimiento representadas a través de un lenguaje específico denominado *Form Definition Language* (FDL). Este lenguaje parte de la premisa básica que tanto la estructura geométrica como la estructura lógica de un documento pueden ser descritas en términos de un conjunto de áreas rectangulares. Dichas áreas producidas en la estructura

[160] Yuan Y. Tang y sus colaboradores citan la obra de S. Tsujimoto y H. Asada. *Understanding multi-articled documents*, para dar a conocer un proyecto piloto instalado en una estación de trabajo SUN-3 que, utilizando 106 documentos distintos obtenidos de revistas, periódicos, libros, manuales, artículos científicos, etc., y

geométrica son analizadas por las reglas marcadas en el lenguaje FDL para obtener la correspondencia lógica[161]. Este método se ha empleado con resultados satisfactorios en el tratamiento de la documentación generada por las Naciones Unidas.

Existen otras técnicas de análisis procedentes de otros campos diversos, como el de la programación informática o el de la gestión y administración de los procesos de trabajo dentro de las organizaciones. En este último caso, existen diversos proyectos que han aplicado el concepto de la modelización de los procesos de trabajo (*modelling work processes*), para la mejora de éstos dentro de las organizaciones, a tareas relacionadas con la generación y análisis de los documentos gestionados dentro de estas instituciones[162]. Muchas de las técnicas que se han empleado dentro de este campo de la gestión y administración de empresas proceden del campo de la computación y programación informática, dado que estas técnicas detallan clara y minuciosamente el proceso de producción de los programadores informáticos para el desarrollo del *software*. Algunas de ellas, entre las que se pueden destacar los *procedural programming languages*, los *data flow diagrams*, las *Petri nets*, los *state-transition diagrams*, o los *control flow languages*, suelen plasmar las etapas o fases de producción a través de un conjunto de notaciones gráficas denominado *Unified Modelling Language* (UML), basado en el lenguaje común de técnicas de modelado con orientación a objetos.

Pero todos estos tipos de métodos no están al alcance de cualquier organización dados sus altos costes de implementación y fuerte especialización, no siendo, además, sus resultados todo lo satisfactorios que cabría desear. Es por todo ello por lo que las más de

---

aplicando estas técnicas de transformación, obtuvo como resultado que sólo 12 documentos no fueran interpretados correctamente. Para una más amplia información véase Y. Y. Tang [*et al.*]. *Op cit.*, p. 16.

[161] Este lenguaje utiliza reglas del tipo "si dentro de un determinado documento que tiene tanto de ancho y tanto de alto se encuentra una caja que mide tanto de ancho y de alto y le sigue otra con tales y cuales medidas, entonces esta última caja corresponderá a tal objeto lógico".

[162] A. Salminen. "Putting documents into their..." *Op. cit.*, p. 628.

las veces suelen emplearse **métodos intuitivos**[163] basados en la observación directa de los documentos y la comparación entre ellos. En este método se hace uso de la experiencia de los autores de los documentos en la confección cotidiana de los mismos, así como en otros factores relacionados como pueden ser lo impuesto por la corporación tanto para las normas de redacción como para el estilo de presentación de los mismos, la disposición, en su caso, de normativa legal que regule los tipos documentales que deben existir y los elementos obligatorios que se han de integrar dentro de cada uno de ellos, etc[164]. Este factor es decisivo, por ejemplo, para la transición que se está produciendo dentro de las Administraciones Públicas del documento tradicional al documento electrónico en donde, como acertadamente señala, L. Auñón la definición de las partes sustanciales del documento administrativo (tanto los elementos materiales, como los subjetivos y, principalmente, los elementos formales) han de producir el establecimiento del tan necesario estudio de los tipos documentales que han de conformar el expediente administrativo en formato electrónico[165].

En cualquier caso, el punto de partida casi siempre será la observación detallada y el análisis directo del documento físico para extraer su estructura física o geométrica. Es

[163] Con el término "intuitivo" no se pretende restar importancia y complejidad al proceso de análisis y modelización formal de los documentos que ha de llevarse a cabo para el desarrollo de DTDs SGML, más bien al contrario. Las habilidades y destrezas necesarias para llevar a cabo esta labor en cualquier organización suelen ser las de más alta cualificación profesional. En este apartado de la tesis se han simplificado deliberadamente (por no alargar en exceso este capítulo) las fases que muchos especialistas en la materia señalan como necesarias para la correcta confección de DTDs. En uno de los manuales imprescindibles dentro de este campo se señalan y analizan con detenimiento las cuatro fases de trabajo para la confección de DTDs SGML dentro de cualquier organización: preparación, análisis de las necesidades, modelización y descripción. Para una mayor profundidad en el tema recomendamos la lectura de la obra de Eve Maler, Jeanne El Andaloussi. *Developing SGML DTDs: From Text to Model to Markup.* Upper Saddle River, NJ: Prentice Hall, 1996.

[164] Por ejemplo, en el caso de la Administración pública española, la Ley 30/1992, de 26 de noviembre, de Régimen Jurídico de las Administraciones Públicas y del Procedimiento Administrativo Común junto con el "Manual de Documentos Administrativos", elaborado por el Ministerio para las Administraciones Públicas, establecen una clara distinción de tipos documentales que se han de emplear en dicha jurisdicción, para los documentos emanados tanto por la Administración pública (certificados, resoluciones, oficios, informes, etc.) como del ciudadano para comunicarse con la administración (instancias, denuncias, alegaciones, recursos, etc.), estableciendo, de igual forma, una estructura interna básica para la composición de los mismos.

recomendable trabajar con una muestra amplia y variada de documentos de cada uno de los posibles tipos documentales que se gestionan en la organización, pues de este modo se podrían ver y recoger las variaciones que se producen entre unos documentos y otros dentro de cada uno de los tipos documentales a describir, siendo contempladas todas estas alteraciones (repeticiones, frecuencias de aparición, omisiones de elementos en determinados casos, etc.) en el modelo final. El resultado de todo ello se suele plasmar de modo sencillo a través de un árbol jerarquizado que muestra la estructura física genérica para cada uno de esos tipos documentales.

La traducción y traslación de la estructura física genérica a su correspondiente estructura lógica genérica no resultará siempre sencilla dado que se han de contemplar las tres funciones más habituales para las que se suelen construir los sistemas documentales basado en SGML (y, por extensión, como se verá posteriormente, los sistemas basados en XML): la generación de bases de datos que faciliten la gestión, recuperación e intercambio de la información; la reutilización de partes de documentos para la generación de nuevos documentos; y la presentación física de los documentos ante un determinado medio de salida. La construcción de la DTD podrá variar enormemente si se desarrolla pensando exclusivamente en alguna de estas funciones o si, por el contrario, se toman en cuenta estas tres funciones, caso éste el más habitual en la mayoría de las organizaciones y el que defenderemos en esta tesis doctoral.

El método más común, directo y sencillo es partir en un primer momento de la traslación directa de los objetos deducidos en la estructura geométrica a la estructura lógica, asociando aquellos objetos que tengan unas mismas características semánticas a un mismo nombre de elemento suficientemente representativo del contenido al que marca. Estos elementos lógicos son los que algunos autores, como es el caso de K. Summers, establecen como propios de la **Estructura Lógica Primaria** (*primary logical structure*), que no es otra que aquella resultante de una traslación directa de los elementos que componen el árbol

[165] Luisa Auñón Manzanares. "Administración Central: del documento tradicional al electrónico. El tipo documental como invariable punto de referencia". *Boletín de la ANABAD*, v. XLV, nº 1, enero-marzo 1995, p. 18.

jerárquico de la estructura física[166]. Tomando el ejemplo anterior del Real Decreto publicado en el BOE, su estructura lógica primaria podría se la siguiente:

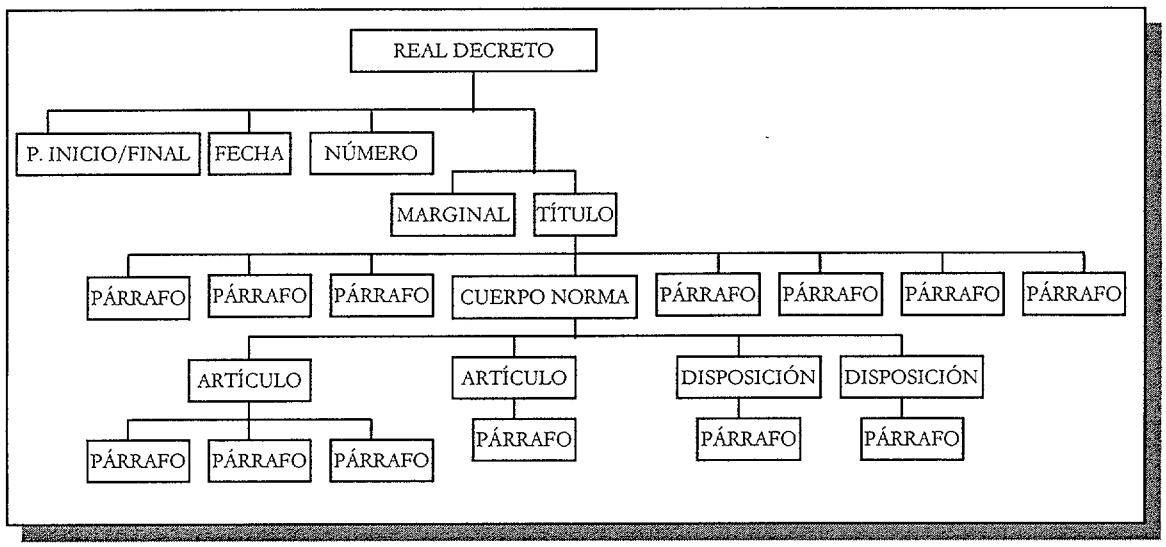


Figura II.10: Estructura lógica primaria de un Real Decreto.

Pero como se observa a simple vista, en esta estructura lógica primaria existen ciertos elementos desagregados, sin una definición clara de jerarquía o pertenencia a un grupo. Es, por tanto, necesario realizar una abstracción de este modelo, añadiendo otros elementos contenedores nuevos cuya función será la de agrupar lógicamente a aquellos elementos que han quedado sueltos y que, por tanto, hacen inconsistente el modelo estructural. Estos nuevos elementos son los que conforman una nueva estructura lógica, denominada **Estructura Lógica Secundaria** (*secondary logical structure*)[167].

Con estos nuevos elementos se perfilan y ordenan de forma más coherente la jerarquía de los elementos dentro del árbol que representa a la estructura lógica del documento, tal y como puede observarse en el siguiente diagrama:

[166] K. Summers. *Op. cit.*, http://www.cs.dartmouth.edu/~samr/DAGS95/Papers/summers.html

[167] *Ibid.*, http://www.cs.dartmouth.edu/~samr/DAGS95/Papers/summers.html

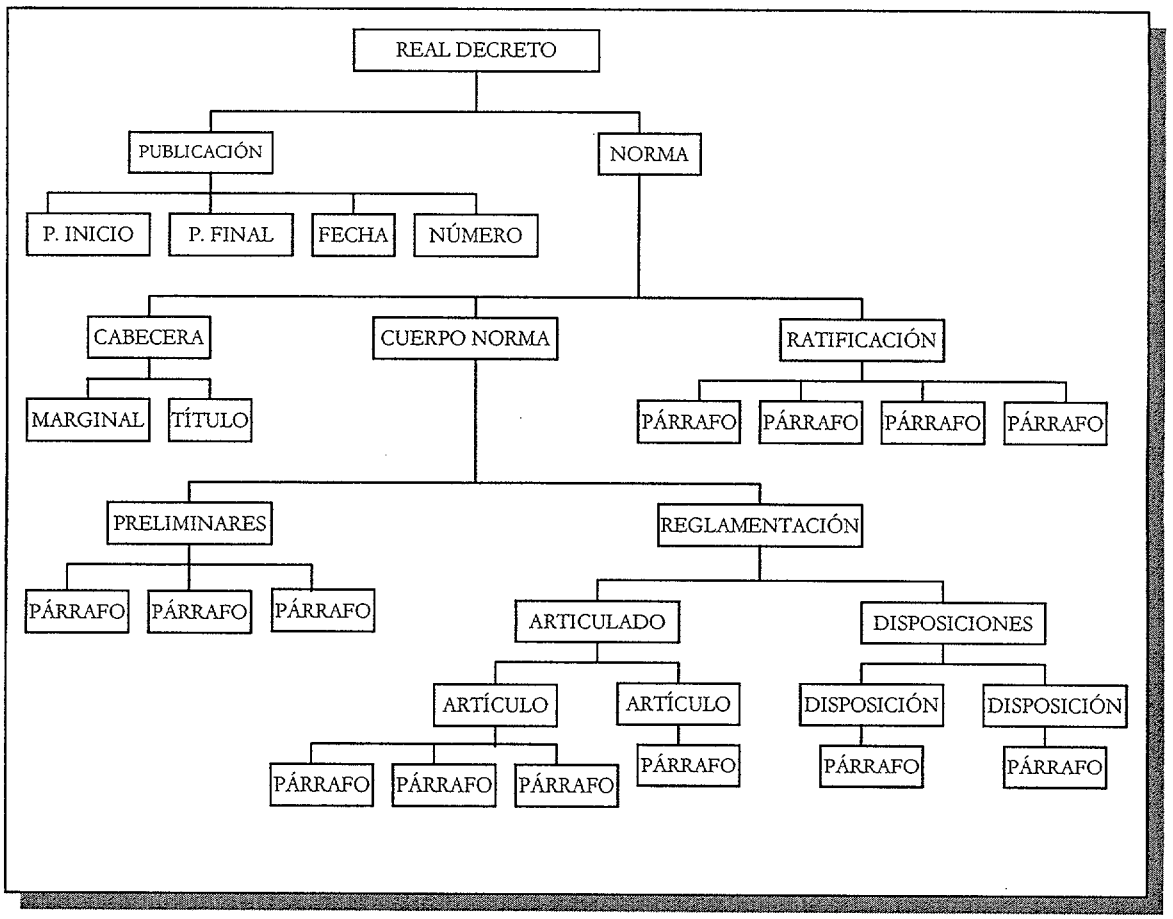


Figura II.11: Estructura lógica secundaria de un Real Decreto.

De este modo, la unión de ambas estructuras lógicas, la primaria y la secundaria, daría como resultado la estructura lógica final para este documento en cuestión. Pero, tal y como se comentó con anterioridad, es posible definir estructuras lógicas teniendo en cuenta la función final para la que ha sido creado el sistema SGML dentro de la organización; esto es, bien la gestión de contenidos a través de bases de datos, bien la reutilización de partes del documento, bien la presentación de los documentos ante un determinado medio de salida, o, las más de las veces, todo ello en conjunto. Por ello, dependiendo de una orientación u otra, la estructura lógica final resultante (y por extensión la futura definición del tipo de documento o DTD) puede ser distinta.

Siguiendo los postulados establecidos por K. Summers, una **estructura lógica orientada al contenido** supondría, en gran medida, que los elementos definidos o por definir deben contener aquella información de interés susceptible de ser localizada y recuperada dentro de un sistema de almacenamiento de bases de datos. De este modo, si se desea una alta capacidad de recuperación se deben marcar otras partes del documento que no han sido contempladas hasta el momento. Normalmente, orientar una estructura lógica hacia el contenido para su posterior localización y recuperación puede suponer que dicha estructura se asemeje bastante poco a su correspondiente estructura geométrica, dado que la finalidad aquí no sería la de la presentación física del documento. Por ejemplo, dentro de la mayoría de los párrafos de la norma se suelen hacer referencias a normas anteriores, bien como apoyo jurídico del legislador o bien porque van a ser modificadas o derogadas.

La recuperación de esta información puede resultar de gran valor para cualquier jurista dentro de un sistema automatizado de recuperación de normativa legal pues nos permitiría, por ejemplo, obtener cómoda y rápidamente todos aquellos documentos en los que se cita una determinada norma (amén de poder establecer otras capacidades de gran interés, como la de relacionar hipertextualmente todas estas citas entre normas) y conocer en qué lugar exacto de la norma se produce dicha referencia. Desarrollando este ejemplo, ciertos párrafos del documento (tanto de los preliminares como de la reglamentación) podrían contener una o varias de estas referencias.

Por el contrario, si en el modelo de definición del sistema SGML sólo interesa la publicación de los documentos estaremos ante una **estructura lógica orientada a la presentación**, en la que deben aparecer reflejados todos los elementos de tipo textual o gráfico que se desea que se presenten en el medio de salida establecido. En este caso, la estructura geométrica del documento debe ser representada lógicamente de la manera más fiel posible. Así, y tomando como ejemplo el apartado de reglamentación del Real Decreto, se observa que en el documento impreso aparece la palabra "DISPONGO", que es la que da entrada a los diferentes artículos y disposiciones de la norma. En la anterior estructura lógica quedaba reflejado este elemento como elemento lógico o semántico, contenedor de los diferentes elementos hijos que se supeditan a él. Pero dicha palabra no tenía una

plasmación como encabezado que es en la estructura lógica (aunque tan sólo se trate de una única palabra) pero sí, tal y como se representó, en la estructura geométrica. Si se desea que esta palabra aparezca finalmente en el documento que ha de ser impreso o presentado en la pantalla del ordenador, ésta deberá tener su correspondiente elemento asignado dentro de la estructura lógica del documento. Lo mismo se puede decir para el resto de encabezados que existen dentro del articulado y de las disposiciones.

Fusionar en un mismo modelo de representación de la estructura lógica ambas orientaciones no es siempre tarea fácil dado que han de contemplarse todas las necesidades, tanto de recuperación documental como de publicación, que la organización contemple. Es por este hecho, tal y como se dijo con anterioridad, por lo que el desarrollo de la mayor parte de las DTDs que se construyen dentro de las organizaciones conlleva la participación de un equipo multidisciplinar, integrado por miembros de diversas áreas de producción de la organización.

Establecida finalmente la estructura lógica del documento particular analizado es necesario contrastar dicha estructura con las resultantes del análisis de los otros documentos de su misma tipología tomados de la muestra. Esto revela otros aspectos de vital importancia para la construcción final de la DTD: la existencia de otros elementos que no son fijos o regulares en su aparición (por ejemplo, algunos Reales Decretos, al igual que otras normas legales, llevan un apartado de anexos, el articulado, cuando es extenso, suele venir enmarcado dentro de capítulos y éstos, en otros, casos dentro de títulos, las disposiciones pueden ser adicionales, transitorias, derogatorias y finales, etc.), las posibles variaciones que se pueden dar en el orden de aparición de los elementos, la frecuencia de aparición de los mismos, etc., y, por último, algo ciertamente complejo y controvertido, como se expuso en su momento, la inclusión de atributos dentro de ciertos elementos.

Toda esta información extra determinará un modelo de **Estructura Lógica Genérica** del tipo documental que se está analizando, paso previo y absolutamente necesario para la redacción final de la Definición de Tipo de Documento.

Una estructura lógica genérica, y más concretamente el modelo de contenido para cada uno de los elementos integrantes en dicha estructura, resulta difícilmente representable a través de gráficos arborescentes dado que ahora se debe plasmar información más puntual de cada uno de estos modelos de contenido, y aquellos no permiten describir las relaciones que se dan entre los diversos elementos del árbol, o la aparición esporádica de ciertos elementos (elementos no siempre obligatorios en el modelo), o su frecuencia de aparición, etc. Es por todo ello por lo que para representar estructuras lógicas genéricas se prefiera el empleo de diagramas de estructura[168].

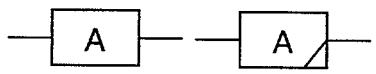
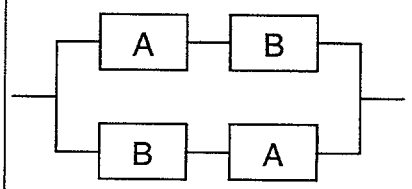
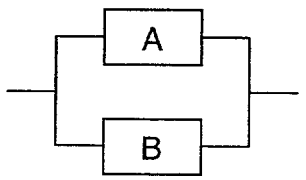
Los **diagramas de estructura** (*structure diagrams*) son un formalismo de representación gráfica utilizado para proporcionar una mayor claridad y un mayor entendimiento en la traslación que se ha de producir entre los resultados del análisis documental y la redacción de la DTD[169]. Los diagramas de estructura se asemejan bastante a los diagramas de flujo pues la notación que se emplea aquí utiliza una secuencia de elementos que van de la izquierda a la derecha y en los cuales se añade gráficamente información relativa a su producción. Existe un diagrama de estructura distinto para cada uno de los modelos de contenido de cada elemento contenedor. De este modo, el conjunto completo de los diagramas de estructura para un determinado tipo documental resultaría de la traslación de cada nodo del árbol jerárquico que representa la estructura lógica genérica siguiendo una determinada representación gráfica para cada uno de ellos. Esta representación gráfica está regida por una serie de reglas de construcción que describen la secuencia de elementos para cada uno de los modelos de contenido, así como las ocurrencias o frecuencias de aparición de cada uno de los elementos contenidos. Así, dentro de estos diagramas encontraremos dos clases distintas de reglas de producción: reglas de conexión (*rules of connection*) y reglas de frecuencia de aparición (*rules of occurrence*).

---

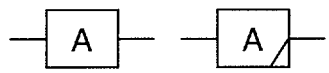
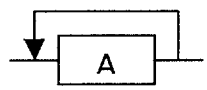
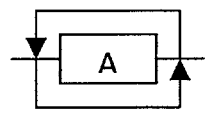
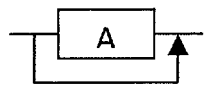
[168] E. van Herwijnen. *Op. cit.*, p. 62.

[169] Rob Stocker. "Module 3 Topic 1: Document Structure / Representation and SGML" [documento HTML]. En: *ITC130 On-Line Publishing.* Bathurst: Charles Sturt University, July 2000. Disponible en http://lorenz.mur.csu.edu.au/itc130/module130/topic32.html (consultado el 24 de octubre de 2000).

Las reglas de conexión expresan la relación existente entre los elementos, y tendrán su posterior redacción a través de los conectadores (la coma, el signo del *ampersand* o la barra vertical) en la definición de la DTD SGML. El siguiente cuadro ilustra las reglas de conexión existentes para esta técnica de representación de estructuras lógicas documentales:

Símbolo	Descripción	Sintaxis en el modelo de contenido	Ejemplo
	Un elemento A simple. Si dicho elemento es terminal (no tiene subelementos) será representado con su esquina inferior derecha marcada (normalmente, en negro).		
,	Conector de secuencia: todos los elementos dentro del modelo de contenido deben aparecen y en el orden que se establece de izquierda a derecha, separados cada uno de ellos por una coma.	SEC (A,B)	
&	Conector AND: todos los elementos dentro del modelo de contenido deben aparecer pero no importa el orden de aparición de los mismos.	AND (A&B)	
|	Conector OR: sólo uno de los elementos dentro del modelo de contenido puede aparecer.	OR (A|B)	

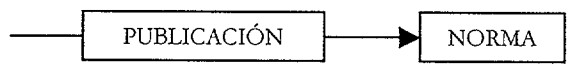
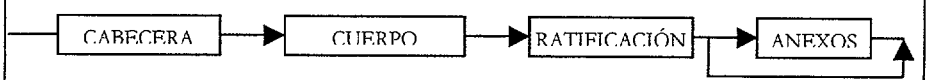
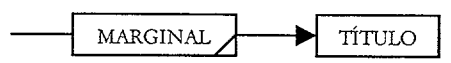
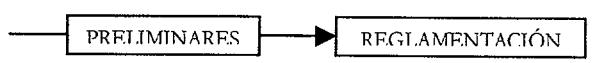
Las reglas de frecuencia de aparición definen el número de veces que se puede suceder un elemento o un conjunto de elementos dentro de cualquier instancia de documento del tipo documental que se está definiendo. El siguiente cuadro ilustra la representación gráfica empleada por esta técnica para cada una de las reglas de ocurrencia:

Símbolo	Descripción	Sintaxis en el modelo de contenido	Ejemplo
	Un elemento A simple que deberá aparecer siempre una única vez.		
+	Indicador de aparición requerido y repetible: el elemento o grupo de elementos puede aparecer una o más veces en la instancia de documento.	MAS (A)+	
*	Indicador de aparición opcional y repetible: el elemento o grupo de elementos puede no aparecer o aparecer tantas veces como sea necesario en la instancia de documento.	REP (A)*	
?	Indicador de aparición opcional: el elemento o grupo de elementos puede no aparecer pero en el caso de hacerlo, lo hará una sola vez.	OPC (A)?	

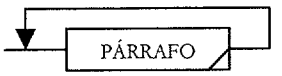
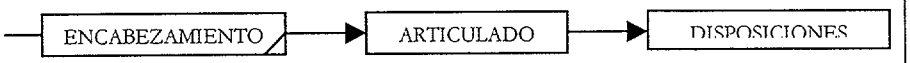
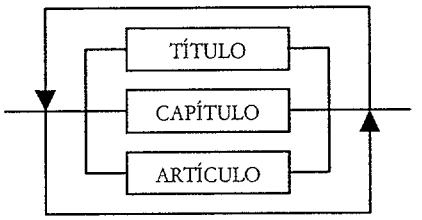
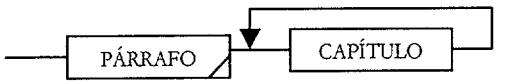
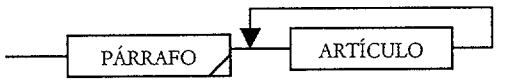
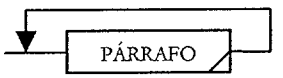
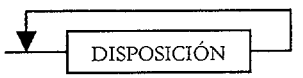
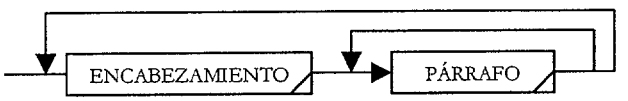
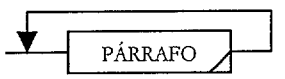
El siguiente paso en el proceso de construcción de la DTD sería, por tanto, trasladar cada nodo contenedor del árbol jerárquico que representa a la estructura lógica genérica (con todos los añadidos y observaciones que se han realizado anteriormente) a sus correspondientes diagramas de estructura. Tomando el ejemplo que se ha venido

desarrollando, el caso de los Reales Decretos, podríamos obtener los siguientes diagramas de estructura para dicho tipo documental[170]:

La siguiente tabla recoge los diagrama de estructura para los diversos nodos o elementos, comenzando por el elemento de documento, REAL DECRETO:

NODO	DIAGRAMA
REAL DECRETO	
PUBLICACIÓN	
NORMA	
CABECERA	
TÍTULO	
CUERPO NORMA	

[170] Se trata de un pequeño ejemplo ilustrativo y, por tanto, desarrollado de modo simple (faltarían otros elementos importantes por definir) con la única finalidad de describir claramente los procesos que se han de desencadenar y las técnicas que se han de emplear para formar correctamente definiciones SGML de tipos documentales. Por tanto, dicho ejemplo debe ser tomado como tal y no como algo cerrado como propuesta de definición de DTD para los Reales Decretos. Otras tesis doctorales que se están produciendo en el Departamento de Biblioteconomía y Documentación de la Universidad Carlos III de Madrid tratarán con mayor profundidad y rigor el desarrollo de DTDs XML para documentos legislativos.

PRELIMINARES	
REGLAMENTACIÓN	
ARTICULADO	
TÍTULO	
CAPÍTULO	
ARTÍCULO	
DISPOSICIONES	
DISPOSICIÓN	
RATIFICACIÓN	

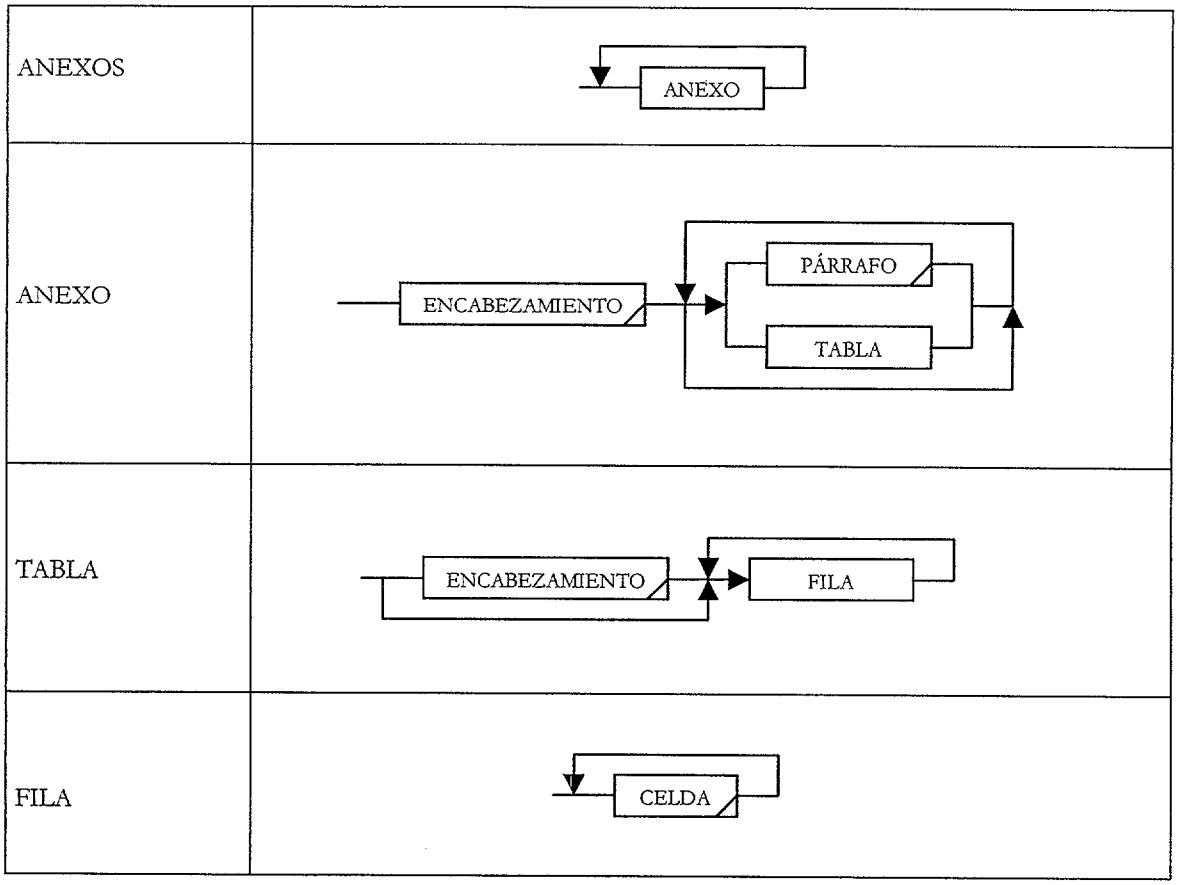
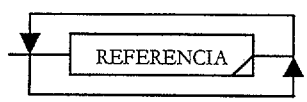


Figura II.12: Tabla de diagramas de los diferentes nodos establecidos para un Real Decreto.

Por último, recordar que anteriormente se comentó la posibilidad de añadir nuevos elementos de interés en el caso de orientar la estructura lógica hacia el contenido. En dicho caso se consideraba oportuno, a efectos de localización, recuperación y posible navegación hipertextual añadir un elemento que referenciase las normas legales citadas en los Reales Decretos. Este nuevo elemento, al cual se le puede denominar "REFERENCIA", sería un elemento hijo del elemento "PÁRRAFO". Así, si se opta por establecer esta idea el gráfico que representa al párrafo no tendría la marca en la esquina inferior derecha (no sería un elemento terminal) y, consiguientemente, habría que añadir un nuevo diagrama de estructura para el nodo "PÁRRAFO", quedando del siguiente modo:



Obtenidos todos los diagramas de esquema del modelo documental tratado tan sólo resta traducirlos a la sintaxis SGML correcta para definir su DTD, teniendo en cuenta, además, la inclusión de atributos para algunos elementos (no definidos en estos gráficos). Si no se ha realizado en una fase anterior (como es nuestro caso), aquí debemos establecer los nombres definitivos de elementos (identificadores genéricos) y atributos, teniendo en cuenta lo establecido por el lenguaje SGML al respecto.

Una posible Definición de Tipo de Documento (DTD) para el caso de los Reales Decretos, ejemplo que se ha seguido para ilustrar esta exposición, podría quedar definitivamente de la siguiente forma:

```
<!ENTITY % P "PARRAFO">
      <!ELEMENT PARRAFO - O (REF)*>
            <!ATTLIST PARRAFO tipo (normal | titulo | numerado) "normal">
      <!ELEMENT REF - - (#PCDATA)>
<!ENTITY % E "ENCABEZAMIENTO">
      <!ELEMENT ENCABEZAMIENTO - O (#PCDATA)>

<!ELEMENT REAL_DECRETO - - (PUBLICACION, NORMA)>
      <!ELEMENT PUBLICACION - - (P_INICIO, P_FINAL, FECHA, NUMERO)>
            <!ELEMENT P_INICIO - O (#PCDATA)>
            <!ELEMENT P_FINAL - O (#PCDATA)>
            <!ELEMENT FECHA - O (#PCDATA)>
                  <!ATTLIST FECHA entrada NUMBER  #REQUIRED>
            <!ELEMENT NUMERO - O (#PCDATA)>
      <!ELEMENT NORMA -  - (CABECERA, CUERPO_NORMA, RATIFICACION, ANEXOS?)>
            <!ELEMENT CABECERA - - (MARGINAL, TITULO)>
                  <!ELEMENT MARGINAL - O EMPTY>
                        <!ATTLIST MARGINAL entrada NUMBER #REQUIRED>
                  <!ELEMENT TITULO - O (%P;)>
            <!ELEMENT CUERPO_NORMA - - (PRELIMINARES, REGLAMENTACION)>
                  <!ELEMENT PRELIMINARES - O (%P;)+>
                  <!ELEMENT REGLAMENTACION - O (%E;, ARTICULADO,
DISPOSICIONES)>
                        <!ELEMENT ARTICULADO - O (TITULO | CAPITULO |
ARTICULO)+>
                              <!ELEMENT TITULO - O (%P, CAPITULO+)>
                              <!ELEMENT CAPITULO - O (%P, ARTICULO+)>
                              <!ELEMENT ARTICULO - O (%P;)+>
                        <!ELEMENT DISPOSICIONES - O (DISPOSICION)+>
                              <!ELEMENT DISPOSICION - O (%E;, %P;+)+>
                                    <!ATTLIST DISPOSICION tipo
(adicional |
                                          transitoria | derogatoria | final)
#REQUIRED>
            <!ELEMENT RATIFICACION - - (%P;)+>
            <!ELEMENT ANEXOS - - (ANEXO)+>
                  <!ELEMENT ANEXO - O (%E;, (%P; | TABLA)*)>
                        <!ELEMENT TABLA - O (%E;?, FILA+)>
                              <!ELEMENT FILA - O (CELDA)+>
                                    <!ELEMENT CELDA - O (#PCDATA)>
```

Figura II.13: Definición de Tipo Documental SGML de un Real Decreto.

Establecida la DTD para el tipo documental de los Reales Decretos, cada vez que se desee crear una instancia de documento SGML correspondiente a un Real Decreto, nuevo documento deberá llevar, además de la correspondiente declaración de este tipo documental y la ruta en la cual se encuentra ubicado el fichero externo de esta DTD, un marcado del texto ajustado a lo declarado para este tipo documental.