UNIVERSIDAD DE SEVILLA

ESCUELA SUPERIOR DE INGENIEROS

INGENIERÍA DE TELECOMUNICACIONES

PROYECTO FIN DE CARRERA

CARACTERÍSTICAS DE MINIMIZACIÓN EN LOS LENGUAJES DE MARCADO: MEDIDAS DE LA REDUCCIÓN DE DOCUMENTOS

Ana María Rojas Pacheco
Tutor: Antonio Jesús Sierra Collado
Departamento: Ingeniería Telemática

Agradecimientos

A mi marido, Jesús, y a mi familia, por su ayuda, comprensión, dedicación y cariño.

A mi tutor, Antonio Sierra, por su guía y su paciencia.

Índice general

Índice de figuras

Capítulo 1

Introducción al proyecto

1.1. Motivación y objetivos

El lenguaje SGML es un estándar internacional para definir formas de codificar textos electrónicos publicado en 1986. SGML tiene una gran potencia y flexibilidad, pero también es complejo. Desde su nacimiento han surgido distintas versiones reducidas que lo simplifican, dejando fuera muchas de sus posibilidades, entre ellas las características de minimización.

Las características de minimización disminuyen enormemente la cantidad de etiquetas necesarias para marcar un documento. Son muy útiles, por tanto, para escribir documentación a mano y, además, su aplicación permite obtener archivos de menor tamaño y, en muchos casos, más fáciles de leer. No obstante, cuando se plantea la necesidad simplificar el lenguaje SGML y crear XML para usarlo en la web, estas características son eliminadas. Se consideró que la complejidad que aportaban estas características no estaba justificada con sus beneficios. El menor esfuerzo para escribir documentos a mano dejó de ser una ventaja con los editores que ayudan al usuario a marcarlos y el ahorro de un pequeño porcentaje de memoria no parece importante hoy en día. Sin embargo, a pesar de que finalmente estas características fuesen eliminadas, durante el desarrollo de las especificación de XML se contempló la posibilidad de mantener algunas de ellas (como eliminar los identificadores en las etiquetas de fin de elemento, eliminar las comillas que deben encerrar el valor de los atributos u omitir los atributos cuando tienen valores por defecto [46][47]). Años después, en el desarrollo de una nueva simplificación de XML, MicroXML, aún hay autores que plantean mirar atrás y rescatar algunas de las características de SGML eliminadas en XML, entre ellas algunas características de minimización (por ejemplo, se propuso eliminar el carácter TAGC si el contenido del elemento sólo consiste en otros elementos y eliminar las comillas que encierran los valores de atributo si no contienen espacios en blanco [7]), por ello podría ser útil volver a estudiar la forma de utilizarlas y las ventajas que pueden ofrecer.

El objetivo de este proyecto es el de medir la reducción que experimenta un documento al hacer uso de las características de minimización del lenguaje SGML. Se pretende comparar la complejidad que aportan dichas características frente al ahorro de memoria que pueden conseguir y comprobar si sus ventajas no son tan despreciables como para

dejar de considerar algunas de ellas en nuevos lenguajes de marcado, como MicroXML, que en la fecha de realización de este proyecto sigue en fase de desarrollo.

Para llegar a dicho objetivo se pretende estudiar previamente el lenguaje SGML, así como los lenguajes XML y MicroXML, y sus diferencias con el primero. Una vez conocidos estos lenguajes de marcado, se estudiarán las características de minimización y se comprobará hasta dónde puede reducirse un documento sin que un analizador confunda su estructura. Será necesario instalar un analizador, escoger un ejemplo sencillo al que puedan aplicarse una tras otra todas las minimizaciones y analizar la problemática a la hora de utilizarlas. A continuación se aplicarán las técnicas de minimización al documento elegido para realizar las medidas.

A la hora de cuantificar la reducción que puede llegar a obtenerse en el documento, se decide comparar el número de caracteres correspondientes a etiquetas que hay en el documento completo y en el mismo documento después de aplicarle distintos tipos de minimización. Para ayudar en dicha tarea se pretende escribir un programa en lenguaje Java que acepte una carpeta con un archivo completo y varios minimizados y devuelva el tanto por ciento de reducción conseguido en cada uno.

Los tipos de documentos que más se beneficiarían de este tipo de reducción, por tanto de los más afectados con la eliminación de estas características y uno de los motivos del debate sobre su inclusión en XML, son los que almacenan información de bases de datos. Las etiquetas representan en ellos un porcentaje muy elevado del tamaño del archivo, por ello se ha escogido para realizar las medidas un fragmento de una base de datos. El documento consiste en un archivo con cincuenta registros extraídos de una gran base de datos creada con propósitos educativos por Gio Wiederhold [39].

A la hora de analizar los resultados se desea comprobar, por un lado, las reducciones obtenidas con cada tipo de minimización y con la combinación más eficaz de las mismas y por otro, la dependencia de dicha reducción con el número de registros considerados.

1.2. Estructura del proyecto

El proyecto consta de cuatro capítulos además de este primer capítulo introductorio donde se han visto la motivación y los objetivos del proyecto.

En el segundo capítulo, relativo a los lenguajes de marcado, se estudia SGML, el lenguaje generalizado estándar para el marcado de documentos, XML, una versión reducida de SGML, y MicroXML, que podría ser el futuro de los lenguajes de marcado. El capítulo está dividido en cuatro secciones:

- La Sección 2.1 comienza introduciendo el concepto de lenguaje de marcado y su evolución. Se pretende, con ello, tener una visión general sobre la utilidad de los lenguajes de marcado.
- En la Sección 2.2 se estudia el lenguaje SGML, un estándar internacional para definir formas de codificar textos electrónicos. Se detallan los componentes que debe tener un documento SGML y la utilidad de cada uno de ellos. Se parte de un ejemplo, en el que se analiza la forma de marcar un documento y se insiste especialmente en la sintaxis de la DTD.

- A continuación se pasa a describir XML (Sección 2.3), el lenguaje extensible de marcas, un lenguaje de marcado que surge como subconjunto de SGML en un intento de simplificarlo. Se estudian los conceptos básicos sobre la sintaxis de un documento XML y los motivos que justificaron su desarrollo. Uno de los objetivos de esta sección es comprender las diferencias entre XML y SGML. Con ese fin, se parte del ejemplo visto para SGML y se modifica hasta hacerlo válido en XML.
- Por último, en la Sección 2.4 se trata el lenguaje MicroXML, una versión reducida de XML que pretende llegar, fundamentalmente, a aquellos usuarios que no han utilizado XML debido a su complejidad. Se muestran los objetivos con los que se ha diseñado el lenguaje y se repasan los conceptos de elemento y atributo. Además, se analiza la especificación propuesta para MicroXML en octubre de 2012 y se prueba uno de los analizadores disponibles actualmente.

El tercer capítulo se centra en las características de minimización del lenguaje SGML y está dividido en tres secciones:

- La primera sección del capítulo describe la instalación de dos programas, un analizador, OpenSP, y un editor, TextPad, que se han utilizado para trabajar con los ejemplos y los archivos extraídos de la base de datos. El analizador se usa para comprobar la validez de los archivos minimizados y el editor ayuda a trabajar con documentos que contienen etiquetas.
- En la Sección 3.2 además de estudiar la forma en que deben aplicarse las minimizaciones, se utilizan en un ejemplo sencillo y se comprueban, con el uso del analizador, algunos fallos que pueden cometerse en su utilización. La importancia de este punto radica en entender la aplicación correcta de las técnicas antes de pasar a aplicarlas a un archivo mayor y más complejo, como el de la siguiente sección.
- En la Sección 3.3 se aplican las características de minimización a un archivo extraído de una base de datos. El objetivo último es reducir el tamaño del documento todo lo posible, pero antes, con el fin de poder hacer una comparativa entre las distintas técnicas de minimización se ha optado por aplicar, una a una, cada posibilidad a los primeros cinco registros. El documento de menor tamaño no tiene porqué ser el mejor, deben tenerse en cuenta otros aspectos como la facilidad a la hora de aplicar cada técnica, si complica o no la lectura del documento o la complejidad que puede suponer para las aplicaciones que manejen los archivos. Se muestran los archivos correspondientes a los cinco registros minimizados con cada técnica y un archivo con cincuenta registros minimizados combinando todas las posibilidades.

En el cuarto capítulo se busca obtener resultados midiendo la reducción de cada uno de los archivos generados en el capítulo anterior. Para ello se cuenta el número de caracteres totales de cada archivo y los caracteres correspondientes a etiquetas con la ayuda de un programa escrito en lenguaje Java para tal fin y que se incluye en la Sección 4.1. En la sección 4.2 se muestran los resultados obtenidos en forma de gráficas, de forma que se facilita la comprensión y comparación de los mismos.

En el quinto y último capítulo se exponen las conclusiones del proyecto y posibles líneas futuras.

Capítulo 2

Los lenguajes de marcado

En este capítulo se estudian los lenguajes de marcado. Se empieza introduciendo el concepto de lenguaje de marcado y la evolución que han experimentado desde GML, el lenguaje de marcado generalizado, hasta MicroXML, una simplificación de XML que se plantea como alternativa de futuro. A continuación se describen los lenguajes de marcado SGML, XML y MicroXML. Para cada uno de estos lenguajes se incluye una introducción, con sus conceptos básicos, una explicación de la estructura que deben tener los documentos en cada lenguaje, otros aspectos que deben tenerse en cuenta y, en el caso de SGML y XML, se detalla la forma de desarrollar sus DTDs.

2.1. Introducción a los lenguajes de marcado

Un lenguaje de marcado es un conjunto de reglas que indican cómo añadir información adicional a un documento. Es información que no forma parte del contenido inicial del documento, se añade para aportar datos sobre su estructura, su presentación o sobre la forma de manejarlo. Para poder incluir estos datos se usan etiquetas (o marcas). El lenguaje de marcado muestra cómo deben ser las etiquetas, dónde pueden colocarse y cómo se diferencian del resto del texto.

En general, a la hora de marcar un documento, lo primero es analizar la información y reconocer las distintas partes por las que está compuesta, cada una será un elemento. Después se determina el formato o la estructura que debe tener cada elemento y se añaden las etiquetas a la información.

2.1.1. Tipos básicos de lenguajes de marcado

Se suelen distinguir dos tipos básicos de lenguajes de marcado: Los lenguajes de marcado de procedimiento y los lenguajes de marcado descriptivo.

- Los lenguajes de marcado de procedimiento se centran en la forma de dar instrucciones sobre cómo procesar los datos, es decir, las marcas se utilizan para indicar las operaciones que se van a aplicar a cada elemento. Por ejemplo, con ellas puede especificarse que el título de un capítulo del documento debe tener un tamaño de letra determinado. El lenguaje TEX se englobaría en esta categoría.

- Los lenguajes de marcado descriptivo, en cambio, pretenden que las etiquetas describan la naturaleza de los componentes del documento. Por ejemplo, puede delimitarse qué parte del documento conforma el título. Después será otro sistema el que se encargue de asociar cada componente con unas instrucciones de proceso según las necesidades concretas. SGML es un lenguaje de marcado descriptivo.

Los lenguajes de marcado de procedimiento tienen una gran desventaja respecto a los descriptivos: se pierde información. Si lo único que diferencia a un título del resto del documento es su formato, y se tiene cualquier otro elemento con ese mismo formato, ya no puede saberse qué es un título y qué no lo es. Además, son poco flexibles, si se quiere almacenar un mismo documento con distintos formatos, habría que marcar de nuevo todo el documento, mientras que con un marcado descriptivo sólo haría falta una asociación de cada elemento con su nuevo formato.

2.1.2. Evolución

Charles Goldfarb es considerado uno de los padres de los lenguajes de marcado. Desarrolló, junto con E.J. Mosher y R.A. Lorie, el lenguaje de marcado generalizado (GML). **GML** aplicaba los principios del marcado descriptivo. Además añadió el concepto de un tipo de documento formalmente definido con una estructura de elementos anidados [18]. Goldfarb, continuó añadiendo conceptos que darían lugar a SGML, que finalmente, fue publicado como estándar internacional en 1986.

SGML (Standard Generalized Markup Language) no es estrictamente un lenguaje de marcado, sino una forma de definir uno, un metalenguaje. SGML especifica una forma de precisar, para los documentos de un determinado tipo, qué etiquetas pueden usarse y cómo deben ser usadas. No son etiquetas fijas, pueden usarse etiquetas distintas para cada tipo de documento que se necesite. Por ejemplo, con SGML, pueden definirse etiquetas que describan los elementos que debe tener un informe (como un título o varias secciones), se debe concretar también cómo se organizan estos elementos y después se utilizan las etiquetas definidas para marcar la información. Todo ello, la definición de las etiquetas y la información con sus marcas, forman parte del documento SGML. SGML tiene una gran potencia y flexibilidad, pero también es complejo. Con el nacimiento de la *World Wide Web*, en 1989, se creó y se popularizó enormemente un lenguaje de marcado mucho más simple: HTML.

HTML (HyperText Markup Language) es considerado una aplicación de SGML, sin embargo no nació como tal. Tim Berners-Lee pretendía desarrollar un lenguaje para marcar textos, de forma que pudiesen tener enlaces a otros documentos y transmitirlos de forma segura por internet. Para delimitar las etiquetas, Berners-Lee usó los mismos caracteres que se usan en SGML, incluso algunas etiquetas coincidían por completo, pero había grandes diferencias en los conceptos básicos [1]. Fue en la versión 2.0 de HTML cuando se decidió definirlo como una aplicación de SGML.

HTML tiene un número fijo de etiquetas, que ha ido creciendo según avanzan sus versiones. En principio estaba más orientado a definir la presentación del documento que su estructura, con el tiempo se ha intentado eliminar las etiquetas relativas al formato y aumentar las etiquetas estructurales.

La necesidad de usar documentos con estructuras más ricas en la Web, llevó al desarrollo de otro lenguaje de marcado en 1996. Hasta ese momento se contaba con HTML, que sólo permitía estructuras prefijadas, y con SGML, que sí permitía una gran libertad a la hora de definirlas, pero se consideró que era demasiado complejo para el tipo de información que suele usarse en la Web. Se optó por definir un subconjunto de SGML, XML.

Para desarrollar **XML** (Extensible Markup Language) se eliminaron algunas de las características que daban flexibilidad a SGML, se trataba de simplificar, pero un documento XML debe ser compatible con SGML, es decir, una aplicación SGML debería poder entender un documento XML.

XML intentaba superar las deficiencias de HTML, pero no lo ha sustituido. Se sigue trabajando para modificar y extender las definiciones de HTML, y que así puedan adaptarse a las nuevas necesidades de la Web. Sin embargo, en un intento de combinar ambas tecnologías y sustituir a HTML, en el año 2000 surgió XHTML. **XHTML** (eXtensible Hypertext Markup Language) es una definición de HTML 4 basada en XML.

XHTML es una versión más estricta y limpia de HTML [43], pero no ha tenido el éxito previsto. En 2010 se cancelaron los trabajos para desarrollar XHTML 2.0 con el objetivo de acelerar el progreso de HTML 5 [49].

El futuro de XML sigue sin estar claro para algunos autores. A finales de 2010, se barajaban tres alternativas [5]: **XML 2.0, XML.next** y **MicroXML**. XML 2.0 y XML.next se planteaban como posibles sustitutos, el primero mantendría la compatibilidad con XML, pero el segundo no se diseñaría pensando en tal fin. MicroXML, en cambio, no intentaría sustituir a XML, sería un subconjunto de XML que podría usarse en los casos donde XML resulte demasiado complejo.

En 2011 se publicaron los primeros borradores de MicroXML y en 2012 se crea un grupo de trabajo liderado por James Clark, John Cowan y Uche Ogbuji, que publica su primera especificación en octubre de ese mismo año.

2.2. SGML

En esta sección del capítulo se hablará sobre el lenguaje SGML (**S**tandard **G**eneralized **M**arkup **L**anguage), un estándar internacional para definir formas de codificar textos electrónicos. En la introducción al lenguaje se verán, de forma básica, los componentes que debe tener un documento SGML y la utilidad de cada uno de ellos. Se hablará más detalladamente de todo ello en los siguientes apartados, insistiendo especialmente en la sintaxis de la DTD (**D**efinición de **T**ipo de **D**ocumento).

2.2.1. Introducción al lenguaje SGML

SGML es un metalenguaje, un lenguaje que permite definir otros lenguajes. Su objetivo es el que se ha visto para los lenguajes de marcado: encontrar una forma de añadir información a un documento, que aporte datos sobre el propio documento. Para eso se utilizan etiquetas o marcas. Con SGML pueden definirse las etiquetas necesarias

para describir las distintas partes del documento y, posteriormente, marcar con ellas la información inicial del mismo. Más formalmente, con SGML pueden especificarse [1]:

- La estructura que podrá tener un documento de un tipo determinado.
- Los caracteres que podrán usarse para marcar la información inicial.
- Los conjuntos de texto que podrán usarse más de una vez sin necesidad de repetirlos.
- La información almacenada fuera del documento que podrá ser incorporada posteriormente.
- Las técnicas que se podrán usar para marcar el documento de forma más eficiente, como la minimización.
- La forma en que será procesado el documento.

Para conseguir este objetivo, son necesarias las tres partes que componen un documento SGML: La declaración SGML, el prólogo y la instacia de documento. La **declaración SGML** aporta información necesaria para poder interpretar el documento (por ejemplo, el conjunto de caracteres usados), el **prólogo** incluye la definición de tipo de documento (DTD), que es la que define las marcas y, con ellas, la estructura que puede tener la información inicial para ajustarse a este tipo de documento; y, finalmente, **la instacia de documento**, que contiene los datos de usuario ya etiquetados.

Si se parte de una cierta información que se quiere almacenar en forma de documento SGML, habrá que analizarla, y así determinar las distintas partes que la componen. Cada una será un elemento, y a cada uno de estos elementos, se le asigna un identificador genérico, que debe ser un nombre que describa al elemento. Si se analiza, por ejemplo, la información de la Figura 2.1, se ve que tiene un título, una primera sección (que tiene su propio título y contiene un párrafo), y un apartado de bibliografía.

```
El lenguaje SGML
Sección 1: Introducción
Un documento SGML tiene tres partes: La declaración SGML, el prólogo y la
instacia de documento...
Bibliografía:
ISO 8879. Information processing -- Text and office systems - Standard
Generalized Markup Language(SGML),1986
```

Figura 2.1: Ejemplo de información sin marcado.

Una forma de marcar esta información sería la de la Figura 2.2. En ella se ve cómo se ha dividido la información en elementos.

```
<informe>
<titulo>SGML</titulo>
<seccion estado=''borrador''>
<titulo>Introducción</titulo>
<p>Un documento SGML tiene tres partes: La declaración SGML, el prólogo y
la instacia de documento...</p>
</seccion>
<bibliografia>
<p>ISO 8879. Information processing -- Text and office systems - Standard
Generalized Markup Language(SGML),1986</p>
</bibliografia>
</informe>
```

Figura 2.2: Ejemplo de información marcada con SGML.

Cada elemento consta de una etiqueta de inicio, el contenido del elemento, y una etiqueta de fin. En el elemento `titulo`, por ejemplo, se pueden identificar estas tres partes:

- La etiqueta de inicio: `<titulo>`
- El contenido del elemento: `SGML`
- La etiqueta de fin: `</titulo>`

No hay motivo para que el contenido del elemento sea tan sencillo como el del elemento `titulo`, puede incluir otros elementos formando una estructura jerárquica, como en el caso del elemento `seccion`,que contiene un título y un párrafo. Debe haber un único elemento raíz (aquí sería el elemento `seccion`) y todos los demás estarán contenidos en él. Es importante que cuando un elemento contenga a otros, todos estén bien anidados (no se puede dar por terminado un elemento hasta que no terminen los elementos que contiene). Es decir, no podría encontrarse la etiqueta de fin de sección (`</seccion>`), sin haber dado por acabado cada párrafo.

El elemento `seccion` tiene una etiqueta de inicio distinta a la del resto de elementos. Además del identificador del elemento, contiene información relativa al mismo. En este caso, sobre el estado de la sección, se dice que todavía está en estado de borrador. Es otra forma de introducir información adicional y se conoce como atributo del elemento. También se verá en apartados posteriores.

Con todo lo anterior se ha conseguido etiquetar un documento. A esto habría que añadirle una declaración SGML, que podría ser la declaración por defecto (que se verá en el apartado 2.2.2.1), y una DTD.

La DTD está formada por un conjunto de declaraciones que definen la estructura del documento y las características opcionales que pueden usarse al prepararlo. Conocer de antemano la estructura que van a seguir este tipo de documentos, permite comprobar si todos la siguen correctamente (no es necesario definir una DTD para cada documento, se define una DTD para todos los documentos que van a seguir esa misma estructura, en el ejemplo, los informes). Además de normalizarlos, la DTD permite ahorrar muchas de las etiquetas que se usan. Si el sistema conoce la estructura, puede completar algunas de las

etiquetas que falten basándose en ella (se verá en el capítulo dedicado a minimización). Como ejemplo, en la Figura 2.3 se muestra la DTD que siguen los informes descritos anteriormente.

```
<!ELEMENT informe            - -   (titulo,seccion+,bibliografia)>
<!ELEMENT seccion            - o   (titulo,p+)>
<!ATTLIST seccion estado (borrador|completo) borrador>
<!ELEMENT bibliografia       - o   (p+)>
<!ELEMENT (p|titulo)         - o   (#PCDATA)>
```

Figura 2.3: Ejemplo de DTD.

En el apartado 2.2.3 se verá la sintaxis de la DTD. Por ahora, puede decirse que en ella se especifican los elementos que aparecen en este tipo de documentos y en qué orden lo harán.

De cada línea se puede extraer mucha información:

- En la primera línea se dice que el elemento de tipo informe contendrá elementos de tipo título, sección (el signo '+' indica que pueden ser varias secciones) y bibliografía.
- En la segunda y cuarta líneas se indica que el elemento de tipo sección está formado por un elemento de tipo título y uno o más párrafos, y el elemento bibliografía sólo por párrafos.
- En la tercera línea se define un atributo estado para el elemento sección, que podrá tomar uno de los dos valores indicados (borrador o completo) y cuyo valor por defecto será borrador.
- En la última línea se ve que los elementos de tipo párrafo y título no contendrán más subelementos, sólo los datos del documento original.

Se ha visto una introducción a los elementos del lenguaje SGML. En los siguientes apartados se analizará más formalmente cómo pasar de los datos originales de un documento a un documento SGML. No obstante, se quiere hacer hincapié en las grandes posibilidades que tiene este lenguaje para adaptarlo a las necesidades concretas de cada caso. La sintaxis usada en esta introducción y descrita en los posteriores capítulos es la que se conoce como sintaxis concreta de referencia (Se verá en el apartado 2.2.2.1), pero no es la única posibilidad; puede usarse una variante si es necesario, por ejemplo, cambiar los caracteres que se usan como delimitadores, los tamaños máximos de los identificadores o las palabras que identifican los tipos de declaración. Lo que debe tenerse en cuenta es que, usando la sintaxis concreta de referencia, los documentos que se creen serán válidos en cualquier sistema SGML.

2.2.2. Documento SGML

Tal y como se comentó en la introducción, un documento SGML tiene tres partes: La declaración SGML, el prólogo y la instacia de documento. En este apartado se describirá la sintaxis de cada una de ellas.

2.2.2.1. Declaración SGML

La declaración SGML es lo primero que debe aparecer en el documento. Proporciona información importante para poder interpretar correctamente el documento, como, por ejemplo, el conjunto de caracteres usados o las características adicionales que pueden utilizarse. Puede ser omitida, a no ser que el documento se envíe a otro sistema. En el caso de que la declaración se omita, se asumiría una declaración por defecto, normalmente la de un documento SGML básico. Según la norma ISO 8879 [19], un documento básico SGML (basic SGML document) es el que sigue dicha norma, usa la sintaxis concreta de referencia y las características de minimización SHORTTAG y OMITTAG. También usa la característica SHORTREF gracias a la sintaxis concreta de referencia. Se muestra esta declaración a modo de ejemplo en la Figura 2.4 [1].

```
<!SGML "ISO 8879:1986"
     -- Declaration for typical Basic SGML Document --
CHARSET 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
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 NONE
>
```

Figura 2.4: Declaración para un documento SGML básico.

La declaración comienza con el delimitador mdo (`<!`) y la palabra `SGML` seguida del número y la fecha de la norma ISO que describe SGML, y termina con el delimitador mdc (`>`). Siguiendo el ejemplo, se describirán las distintas partes de la declaración:

- `CHARSET` proporciona información sobre el conjunto de caracteres utilizados. En

concreto, `BASESET` contiene la declaración del conjunto de caracteres, aquí es el `ISO 646`, y `DESCSET` contiene una descripción de la forma en que serán usados (por ejemplo: la primera línea indica que las primeras nueve posiciones, del 0 al 8, no se usarán en el documento).

- `CAPACITY` informa sobre las restricciones de capacidad. Es una limitación en el número de caracteres de marcado de un documento. Se fija el máximo en 35.000 caracteres, sin embargo, la mayoría de los sistemas suelen ignorar esta restricción dando solo un mensaje de aviso al superarlo.
- `SCOPE` es el ámbito de validez de la declaración. Por defecto será el documento completo, pero puede cambiarse su valor a `INSTANCE`, si el conjunto de caracteres especificado solo se usa para marcar el texto y no en las declaraciones.
- `SYNTAX` indica la sintaxis concreta que usa. En este caso es la sintaxis concreta de referencia. Se hablará de ella más adelante.
- `FEATURES` se refiere a las características adicionales que pueden usarse en el documento. Aquí permite el uso de `OMITTAG` y `SHORTTAG.` Se verán en el apartado dedicado a minimización.
- `APPINFO` puede incluir información que sea necesaria para procesar el documento.

Como se ha comentado, un documento básico usa la sintaxis concreta de referencia. Se definen dos tipos de sintaxis para SGML, la abstracta y la concreta. Según la norma ISO 8879 :

- Sintaxis abstracta: Son las reglas que definen cómo se añade el marcado a los datos, sin especificar los caracteres concretos que lo representan.
- Sintaxis concreta: Es la asociación de la sintaxis abstracta con los caracteres concretos que actuarán de delimitadores. Además especifica los caracteres que deben ignorarse (por ser caracteres de control), los caracteres de función, los caracteres que pueden usarse en los identificadores, las palabras reservadas y los caracteres que pueden usarse como referencias cortas. Estos últimos se verán en el apartado 3.2.3.

La sintaxis concreta de referencia es la sintaxis concreta que pretende ser un modelo para las demás y es la que se usa en todas las declaraciones SGML. Si no se indica lo contrario, siempre que aparezcan entre paréntesis los caracteres asociados a un delimitador, se referirán a dicha sintaxis. El conjunto de caracteres definidos como delimitadores por la sintaxis concreta de referencia se incluye como anexo C.2.

Cuando se incluyen los anexos J (Extended Naming Rules) y K (Web SGML Adaptations) a la norma que define a SGML, la declaración SGML sufre algunas modificaciones para permitir, por ejemplo, tener más control al especificar las características opcionales permitidas, o eliminar las restricciones de capacidad (fijando `CAPACITY NONE`). En ese caso la declaración comienza con `"ISO 8879:1986 (www)"` haciendo referencia a la World Wide Web.

2.2.2.2. Prólogo

El prólogo es la porción del documento SGML que contiene la declaración de tipo de documento y, si existe, la definición de procesos de enlace (Link Process Definition). Este apartado se centrará en la declaración de tipo de documento y en la DTD (definición de tipo de documento), que debe estar incluida en la primera.

- **Declaración de tipo de documento**

La declaración de tipo de documento es la sección que contiene la DTD, o bien, dice dónde puede encontrarse.

Comienza con el delimitador mdo (`<!`), seguido de la palabra DOCTYPE, del nombre del tipo de documento que se va a definir y de la DTD (o su localización). Termina con mdc (`>`).

Si se optase por indicar el archivo en el que la aplicación puede encontrar la DTD, podría hacerse con una declaración como la siguiente:

```
<!DOCTYPE informe SYSTEM "informe.dtd">
```

Aquí "informe.dtd" es el nombre de un archivo almacenado en el sistema. Otra posibilidad es indicar que se va a usar una DTD pública. Un ejemplo sería el siguiente:

```
<!DOCTYPE HTML PUBLIC ''-//W3C//DTD HTML 4.01//EN''>
```

Aquí , `''-//W3C//DTD HTML 4.01//EN''`, es un identificador público.

Si, en cambio, se quiere incluir la DTD en el mismo archivo, se usa lo que se conoce como subconjunto de la declaración. El subconjunto de la declaración debería ir correctamente delimitado por dso (`'['`) y dsc (`']'`). En la Figura 2.5 se muestra una declaración de tipo de documento que contiene un subconjunto con la DTD de la Figura 2.3.

```
<!DOCTYPE informe[
<!ELEMENT informe            - -   (titulo,seccion+,bibliografia)>
<!ELEMENT seccion            - o   (titulo,p+)>
<!ATTLIST seccion estado (borrador|completo) borrador>
<!ELEMENT bibliografia       - o   (p+)>
<!ELEMENT (p|titulo)         - o   (#PCDATA)>
]>
```

Figura 2.5: Ejemplo de DTD incluida como subconjunto.

Pueden usarse ambas posibilidades combinadas para ampliar la DTD de un archivo con nuevos elementos definidos en el subconjunto de la declaración.

- **Definición de tipo de documento**

La definición de tipo de documento (DTD) contiene las normas que deben cumplir las instancias de un tipo de documento concreto. Según el anexo B de la norma ISO 8879, la DTD especifica básicamente: los nombres de elemento que podrán usarse (GIs),

los atributos que puede contener cada elemento y su modelo de contenido (el tipo de contenido que puede tener). La DTD no especifica nada sobre las formas en que puede procesarse o formatearse el documento, ni sobre los delimitadores concretos que se usarán en el marcado. En el apartado 2.2.3 se estudiará la estructura de la DTD.

2.2.2.3. Instancia de documento

Es la parte que contiene los datos originales del documento junto con las etiquetas que se emplean para marcarlos. Ya se había visto un ejemplo en la Figura 2.2. En la instacia de documento la información queda estructurada en elementos. El primer elemento en aparecer es el elemento raíz, que contiene a todos los demás (solo puede haber un elemento raíz por documento). Se organizan formando una estructura jerárquica. El orden en que aparecen los elementos, la cantidad de veces que pueden aparecer, sus atributos, y el tipo de datos (o elementos) que pueden contener, debe ajustarse a lo dictado por la DTD.

Se ha comentado que cada elemento comienza con una etiqueta de inicio, y termina con una etiqueta de fin. Esto es así habitualmente pero hay excepciones. Un elemento no tendrá etiqueta de fin si ha sido declarado como vacío o su contenido se proporciona con un atributo de referencia de contenido (se verá en la sección 2.2.3.2), también puede carecer de alguna de las etiquetas o de parte de ellas si se usa alguna minimización.

Las etiquetas están compuestas, básicamente, por el identificador del elemento delimitado por los siguientes caracteres:

- STAGO (start-tag open), "<" , y TAGC (tag close), ">", en el caso de la etiqueta de inicio. Por ejemplo: `<titulo>`, donde `titulo` es el identificador del elemento.
- ETAGO (end-tag open), "</" , y TAGC (tag close), ">", en el caso de la etiqueta de fin. Por ejemplo: `</titulo>`, donde `titulo` es el identificador del elemento.

En la etiqueta de inicio pueden aparecer los atributos del elemento, si los tiene. Los atributos pueden ser usados para aportar información adicional sobre los elementos que los contienen. Según [1], algunos de los usos más comunes de un atributo pueden ser, por ejemplo, identificar el estado de un elemento (`<LIBRO estado=''borrador''>`), identificar una ocurrencia concreta de un elemento (`<FIGURA id=''figura1>''`) o referenciar elementos previamente definidos (`<REFER a=''figura1''>`). A la hora de usarlos en la instancia de documento deben colocarse en la etiqueta de comienzo del elemento. Primero el nombre del atributo y luego el signo '=' seguido del valor entre comillas dobles o simples. Esta estructura puede minimizarse de varias formas usando las características adicionales de SGML. En el ejemplo visto, el elemento `seccion` tiene un atributo cuyo identificador es `estado` y cuyo valor es `borrador`:

```
<seccion estado=''borrador''>
```

En el apartado 2.2.3.2 se verán las posibilidades que ofrecen los atributos.

En cuanto al contenido de los elementos, debe ajustarse a lo que indique la DTD para cada uno de ellos. Según el tipo de contenido que se indique en la declaración de elemento, un elemento podrá contener: datos de usuario, un contenido mixto (puede mezclar datos de usuario con elementos y otro tipo de contenidos que se verán a continuación) o bien

un contenido consistente en otros elementos (podría tener elementos y también alguno de los componentes que se comentan a continuación).

Además de otros elementos y datos de usuario, en el contenido de un elemento pueden aparecer: comentarios, instrucciones de proceso, referencias de caracteres, referencias de entidades generales, secciones marcadas, referencias cortas, etc. La utilidad y la sintaxis de estos componentes se verá a lo largo del capítulo.

2.2.3. Implementación de la DTD

La DTD está formada por un conjunto de declaraciones de marcado que se aplican a todos los documentos de un tipo determinado. Las más importantes son las declaraciones de elementos, las declaraciones de listas de definición de atributos y las declaraciones de entidades. Con ellas se especifican los elementos que podrán usarse en un documento, las relaciones entre ellos, su modelo de contenido, sus atributos (con los valores que pueden tomar) y las entidades que pueden referenciarse. Se verán todas ellas en los siguientes apartados.

2.2.3.1. Declaración de elemento

En la declaración de cada elemento se describen: su nombre, las características de minimización que pueden usarse en el elemento y todas las posibilidades que se tienen en cuanto a su contenido. Los componentes básicos en la declaración de un elemento son los siguientes:

```
<!ELEMENT       nombre      (minimización)      contenido>
```

MDO	Delimitador con el que comienza una declaración (”<!”).
ELEMENT	Palabra reservada que distingue la declaración de un elemento.
nombre	Nombre del elemento concreto que estamos declarando.
minimización	Posibilidad de minimizar sus etiquetas.
contenido	Contenido permitido.
MDC	Delimitador de fin de declaración (”>”).

El nombre que se escoja para el elemento debe empezar por un “name start character”, que básicamente son letras mayúsculas o minúsculas, y puede continuar con letras, números o cualquier otro carácter que permita la sintaxis concreta. También puede contener varios nombres (en caso de declarar varios elementos a la vez). En ese caso, los nombres deben estar entre paréntesis y separados por cualquier conector. Los conectores se ven al hablar del modelo de contenido. Se pueden usar los caracteres ’,’ , ’|’ o ’&’. Por ejemplo:

```
(elemento1|elemento2|elemento3)
```

Respecto a la minimización, se usan dos caracteres como indicadores de la posibilidad de omitir la etiqueta de comienzo, la de fin o ambas. Se hablará de este tipo de minimización al ver la característica OMITTAG. Se comenta, por ahora, que como indicadores se usan dos caracteres separados por un espacio, el primero para indicar si es posible o no omitir la etiqueta de comienzo, y el segundo para la etiqueta de fin. Se usan los caracteres ’o’ (mayúscula o minúscula), para permitir la omisión, y ’-’, para no permitirla.

Como ejemplo de lo anterior, se ha extraído una de las líneas de la Figura 2.3:

```
<!ELEMENT (p|titulo)          - o   (#PCDATA)>
```

En el ejemplo pueden observarse los siguientes puntos:

- Se declaran dos elementos a la vez, `p` (párrafo) y `titulo`.
- Al usarlos, no se podrán omitir sus etiquetas de inicio, pero sí sus etiquetas de fin.
- No pueden contener otros elementos, sólo datos.

Las posibilidades para especificar el tipo de contenido son muchas:

- Pueden usarse las siguientes **palabras reservada**s:
  - `EMPTY`: Para indicar que el elemento no tiene contenido. Si un elemento se declara como vacío, no tendrá etiqueta de fin.
  - `CDATA`: El contenido será considerado como datos de usuario, es decir, no se analizará en busca de caracteres de marcado (salvo los de la etiqueta de fin).
  - `RCDATA`: Similar al anterior, pero aquí permite usar referencias a entidades.
  - `ANY`: El contenido puede ser una mezcla de todos los que se van a ver.
- También puede definirse una estructura con subelementos y datos. Para ello se usan los grupos de modelo de contenido, que pueden estar formados por: elementos, conectores, indicadores de ocurrencia, la palabra reservada "`#PCDATA`" (caracteres de datos que pueden ser analizados en busca de marcado), y otros grupos (pueden anidarse). Cada grupo debe ir entre paréntesis.
  - Los distintos componentes de un grupo se separan con conectores, cada uno con un significado concreto. Dentro de cada grupo sólo puede usarse un tipo de conector. Los **conectores** son:

    `','` Todos los componentes del grupo deben aparecer y deben hacerlo en el orden indicado.

    `'|'` Sólo se escoge uno de los componentes.

    `'&'` Todos los componentes deben aparecer, pero pueden hacerlo en cualquier orden.

    Ejemplo: `(elemento1 | (elemento2 & elemento3 ))`
    El contenido del elemento que se está declarando debe estar formado por un elemento de tipo elemento1, o bien por dos elementos, uno de tipo elemento2 y otro de tipo elemento3 en cualquier orden.
  - Para indicar si los componentes del grupo, o los propios grupos, pueden aparecer cero, una o más veces se usan los indicadores de ocurrencia. Cuando no se especifica indicador de ocurrencia, se entiende que el elemento, o grupo, debe ocurrir una vez. Los **indicadores de ocurrencia** son:

'?' El grupo o elemento es opcional. Si aparece, puede hacerlo una sola vez.

'*' El grupo o elemento puede ocurrir cualquier número de veces (incluso cero).

'+' El grupo o elemento debe ocurrir al menos una vez, pero puede hacerlo más veces.

Ejemplo: `(elemento1,elemento2+,elemento3)`
El contenido del elemento que se está declarando debe estar formado por un elemento de tipo elemento1, uno o más elementos de tipo elemento2 y un elemento de tipo elemento3.

- Cuando la palabra reservada #**PCDATA** aparece sola, con o sin indicadores de ocurrencia, permite cero o más caracteres de datos. Suele aparecer en modelos de contenido que contengan sólo un grupo opcional y repetible. Al poder combinarla con elementos permite tener un contenido mixto. El carácter "#" (llamado "reserved name indicator") sirve para diferenciar una palabra reservada de un identificador genérico.

  Ejemplo: `(#PCDATA|elemento1)*`
  El contenido del elemento estará formado por datos, por elementos de tipo elemento1, o por ambos, cualquier número de veces.

- Cuando se usen modelos de contenido, o bien la palabra reservada ANY, se pueden añadir después **grupos de inclusión o exclusión**, permitiendo especificar excepciones a los modelos de contenido que se definen delante. Estos grupos también deben ir entre paréntesis y, si hay varios componentes pueden separarse con cualquier conector. Los grupos de inclusión se añaden elementos permitidos al modelo de contenido. Deben ir precedidos por el carácter '+'. Los grupos de exclusión suprimen elementos del modelo de contenido. Van precedidos del carácter '-'.

  Por ejemplo:

  ```
  <!ELEMENT informe - - (titulo,seccion+,bibliografia)+(figura)>
  ```

  Es una forma de permitir que las figuras aparezcan en cualquier parte del informe.

Puede verse todo resumido en la Figura 2.6 [23] (los elementos entre corchetes son opcionales):

Figura 2.6: Declaración de elemento.

2.2.3.2. Declaración de la lista de atributos

La declaración de la lista de atributos especifica el conjunto de atributos que pueden usarse con un elemento y los tipos de valores (o los propios valores) que pueden tomar.

La sintaxis general es la siguiente:

```
<!ATTLIST   nombre_de_elemento  lista_definición_atributos>
```

MDO	Delimitador con el que comienza una declaración (”<!”).
ATTLIST	Palabra reservada que distingue la declaración de una lista de atributos.
nombre_de_elemento	Nombre del elemento al que se añaden los atributos.
lista_definición_atributos	Cada atributo de la lista incluye un **nombre** de atributo, un **valor** (que puede ser una lista de posibles valores o un tipo de contenido) y un **valor_por_defecto** (que también puede ser alguna característica del valor del atributo).
MDC	Delimitador de fin de declaración (”>”).

Se habla de lista, porque en la misma declaración se incluyen todos los atributos que puede tener el elemento. La definición de cada atributo está separada de la siguiente por un espacio en blanco o un carácter similar (como tabulador o nueva línea), un comentario o una referencia a entidad (si la entidad comienza con un carácter separador). Por ejemplo:

```
<!ATTLIST seccion estado (borrador|completo) borrador
                  version     NUMBER          ''1''>
```

Para cada atributo de la lista, además de su nombre, se incluye un término relativo a su valor y otro relativo a su valor por defecto.

El término relativo a su valor puede contener una lista de valores entre paréntesis y separados por cualquier conector, como es el caso del atributo `estado` en el ejemplo, o bien, una palabra reservada que indica el tipo de valores que puede incluir, como el atributo `version`, que usa la palabra reservada `NUMBER` para indicar que su valor debe ser numérico. Las palabras reservadas que pueden ser usadas para especificar el tipo de contenido del atributo se muestran en la Figura 2.7.

El término relativo al valor por defecto también puede ser un valor concreto o una palabra reservada (aquí van precedidas por el indicador de nombre reservado, que es el carácter'#'). En el ejemplo anterior, ambos atributos especifican valores concretos como valor por defecto. En el caso del atributo `estado` debe ser uno de los valores de la lista, puesto que son las únicas opciones que tiene, y en el caso del atributo `version` basta con que sea un valor numérico. En el siguiente ejemplo se muestra la otra opción, usar como valor por defecto una palabra reservada.

```
<!ATTLIST imagen archivo ENTITY #REQUIRED>
```

En el ejemplo, el atributo `archivo` debe tener un valor de tipo entidad (se verán en la siguiente sección), y en vez de un valor por defecto, se emplea la palabra reservada `#REQUIRED`, que indica que el atributo es requerido, es decir, que siempre debe aparecer. En la Figura 2.8, pueden verse las palabras reservadas que pueden usarse como valor por defecto.

Palabra reservada	Tipo de contenido del atributo
CDATA	Caracteres de datos válidos para SGML.
ENTITY	Nombre de una entidad ya declarada en el documento.
ENTITIES	Lista de nombres de entidades.
ID	Identificador único que se asociará al elemento.
IDREF	Referencia al identificador de un elemento.
IDREFS	Lista de referencias a identificadores.
NAME	Nombre válido para SGML.
NAMES	Lista de nombres válidos.
NMTOKEN	Nombre que no sigue necesariamente las reglas del carácter de comienzo (El carácter de comienzo de un nombre ,"name start character", debe ser una letra mayúscula o minúscula), y por tanto, puede empezar por un carácter numérico u otro carácter válido para un nombre.
NMTOKENS	Lista de nombres NMTOKEN.
NOTATION	Nombre de notación ya declarada en el documento.
NUMBER	Cadena de caracteres numéricos.
NUMBERS	Lista de cadenas de caracteres numéricos.
NUTOKEN	Nombre que comienza con un carácter numérico.
NUTOKENS	Lista de nombres que comienzan con caracteres numéricos.

Figura 2.7: Palabras reservadas para tipos de valores de atributos.

Palabra reservada	Tipo de contenido del atributo
#FIXED	El atributo tiene un valor fijo.
#REQUIRED	El atributo es requerido (siempre debe aparecer).
#CURRENT	Si no se especifica un valor para el atributo, se usa como valor por defecto el valor de este atributo en el elemento del mismo tipo más cercano (anterior al que nos referimos).
#IMPLIED	El atributo es opcional. Las aplicaciones deben darle su propio valor si es necesario.
#CONREF	Es un atributo de referencia de contenido. El atributo es opcional, pero si se especifica, el elemento no puede tener contenido ni etiqueta de fin. En este caso el valor del atributo normalmente nos dará información para asignar un contenido al elemento, por ejemplo, puede ser una referencia a una entidad o una referencia a un identificador.

Figura 2.8: Palabras reservadas para valores por defecto de atributos.

En la norma ISO 8879 [19] se señalan algunas restricciones relativas al uso de estas palabras reservadas:

- En cada lista de definición de atributos sólo puede declararse un atributo de tipo ID y uno de tipo NOTATION.
- Las palabras reservadas NOTATION y #CONREF no pueden usarse con elementos que hayan sido declarados como vacíos (EMPTY).
- El valor por defecto de un atributo declarado como de tipo ID debe ser #REQUIRED o #IMPLIED.
- Sólo los atributos que hayan sido declarados como CDATA pueden tener como valor por defecto una cadena vacía (“”).

Si se tienen varios elementos con los mismos atributos, pueden compartir la declaración de la lista de atributos. En lugar de un nombre de elemento, se usaría una lista de nombres de elementos entre paréntesis y separados por cualquier conector. Por ejemplo:

```
<!ATTLIST (seccion|bibliografia) estado (borrador|completo) borrador
                                 version     NUMBER          ''1''>
```

Puede verse todo resumido en la Figura 2.9 [23] (los elementos entre corchetes son opcionales). Aquí se declara un solo atributo, pero puede declararse una lista repitiendo el conjunto: nombre de atributo, valor y valor por defecto.

<!ATTLIST	Nombre de elemento Grupo de nombres de elementos	Nombre de atributo	Lista de posibles valores CDATA ENTITY ENTITIES ID IDREF[S] NAME[S] NMTOKEN[S] NOTATION NUMBER[S] NUTOKEN[S]	Valor por defecto #CONREF #CURRENT #FIXED valor #IMPLIED #REQUIRED	>

Figura 2.9: Declaración de atributo.

2.2.3.3. Declaración de entidad

Una entidad no es más que una porción de documento, un conjunto de caracteres o un archivo externo con texto o datos binarios, y sin embargo, las referencias a entidades tienen un buen número de utilidades. En el anexo B de la norma ISO 8879 [19] se destacan las siguientes:

- Pueden usarse para sustituir cadenas largas de caracteres por referencias más cortas.
- Pueden usarse para sustituir caracteres especiales que no pueden introducirse con teclados convencionales.
- Usando referencias para elementos que sean específicos de cada sistema (como los caracteres anteriores, que no pueden introducirse directamente) se facilita el intercambio de documentos, al permitir que sea el sistema receptor el que las resuelva.

- Gracias a las entidades pueden añadirse al documento partes que hayan sido almacenadas por separado.
- Mediante las entidades puede incluirse el resultado de instrucciones de proceso ejecutadas dinámicamente.

La entidad que contiene al documento sgml completo es la entidad documento. La entidad documento (y el resto de entidades) puede tener referencias a otras entidades y, al referenciarlas, son incluidas en el documento. Hay que distinguir entre la declaración de entidad, donde se especifica su contenido, y las referencias a esa entidad, que serán los puntos donde se inserta ese contenido.

- **Entidades generales y paramétricas**

La norma distingue dos tipos de entidades: las generales, que pueden usarse como parte del contenido de un elemento o del valor de un atributo y, por tanto, se referencian en la instancia de documento, y las paramétricas, que contienen parte de una declaración, por lo que serán referenciadas dentro de una de ellas.

La forma de referenciarlas dependerá del tipo de entidad del que se esté hablando:

- Para entidades generales: **&nombre;**
- Para entidades paramétricas: **%nombre;**

ERO	Carácter de apertura de referencia a entidad general (”&”).
PERO	Carácter de apertura de referencia a entidad paramétrica (” %”) .
nombre	Nombre de la entidad.
REFC	Carácter de cierre de entidad (”;”). También se considera fin de entidad un carácter fin de línea u otro carácter (como un espacio) que no sea parte de un nombre válido de entidad.

Para especificar la sintaxis de su declaración también se debe diferenciar entre los dos tipos:

- Para entidades generales:

```
<!ENTITY   nombre    contenido>
```

- Para entidades paramétricas:

```
<!ENTITY % nombre    contenido>
```

A continuación, se muestra el significado de los distintos componentes:

MDO	Delimitador con el que comienza una declaración (“<!”).
ENTITY	Palabra reservada que distingue la declaración de una entidad.
PERO	Carácter de apertura de referencia a entidad paramétrica (“ %”).
nombre	Puede consistir en el nombre de la entidad o en una palabra clave que identifica a la entidad por defecto (#DEFAULT), que será usada cada vez que se encuentra una referencia a entidad que no haya sido declarada.
contenido	El más sencillo es un texto entre comillas, que será directamente el texto que reemplazará a la entidad. Para las paramétricas suele ser una lista de nombres de elementos.
MDC	Delimitador de fin de declaración (“>”).

Ejemplo1: `<!ENTITY entidad1 "entidad general">`

Ejemplo2: `<!ENTITY% entidad2 "#PCDATA | elemento1 | elemento2">`

Se ha comentado que el contenido puede ser simplemente el texto que reemplace a la entidad, pero cuando se habla de entidades generales, también puede ir precedido por un indicador del tipo de datos que contiene. El indicador puede ser una de las siguientes palabras reservadas:

- **CDATA:** La entidad contiene datos de tipo carácter. No debe ser analizada en busca de caracteres de marcado.
- **SDATA:** La entidad contiene datos con caracteres específicos de un sistema.
- **PI:** El texto de la entidad debe ser interpretado como una instrucción de proceso.
- **STARTTAG:** El texto de la entidad debe ser interpretado como una etiqueta de comienzo (como si tuviese los delimitadores adecuados para serlo). Equivaldría a : `<texto>`.
- **ENDTAG:** El texto de la entidad debe ser interpretado como una etiqueta de fin. Equivaldría a : `</texto>`.
- **MS:** El texto de la entidad debe ser interpretado como si tuviese los delimitadores de una sección marcada. Equivaldría a : `<![texto]>`.
- **MD:** El texto de la entidad debe ser interpretado como si fuese una declaración. Equivaldría a : `<!texto>`.

Ejemplo1: `<!ENTITY titulo2 CDATA "La etiqueta <p>" >`. Aunque haya caracteres de marcado, no serán interpretados como tales, pues se ha indicado CDATA como tipo. Así, se consigue que aparezca una etiqueta en el texto sin ser analizada.

Ejemplo2: `<!ENTITY ti STARTTAG ''titulo''>`. La referencia a esta entidad será interpretada como una etiqueta de comienzo del elemento título.

Existe un tipo de declaración especial para declarar una entidad por defecto, que será usada cada vez que se encuentre una referencia a una entidad que no haya sido declarada. La declaración tendría la siguiente forma:

```
<!ENTITY #DEFAULT ''Se ha encontrado una entidad sin declarar''>
```

- **Entidades internas y externas**

Tanto las entidades generales como las paramétricas pueden ser internas o externas según dónde se almacene el contenido de la entidad. Si se almacena en el prólogo se dice que la entidad es interna, si se almacena en un fichero externo, la entidad es externa. Hasta ahora se han visto ejemplos de entidades internas.

En el caso de las externas, el contenido de la entidad está formado por un identificador externo y después, posiblemente, un tipo de entidad (sólo en el caso de que la entidad sea general). El identificador depende de si se añade una localización que es específica del sistema que crea el documento, o si la entidad es declarada pública, con el objetivo de que sea conocida en más de un sistema.

Para las primeras se usará el nombre reservado `SYSTEM`, seguido de una cadena de caracteres que contiene la ruta para llegar hasta el fichero:

```
<!ENTITY  entidad1      SYSTEM     "c:\SGML\ejemplos\entidad1.sgml" >
```

Para las segundas se usará el nombre reservado `PUBLIC` seguido de una cadena de caracteres normalizada que nos da información sobre la entidad. Algunas entidades públicas sólo contienen un conjunto de declaraciones que se añaden al prólogo del documento. La ventaja es que no hace falta enviarlas si el sistema receptor ya tiene acceso a ellas. En principio se llaman públicas porque el sistema receptor las conoce de antemano, pero también hay entidades que pueden llamarse realmente públicas en el sentido de que han sido definidas por una organización autorizada por la ISO.

Si la entidad contiene declaraciones que deben añadirse a la DTD se debe usar una declaración de entidad paramétrica. Por ejemplo:

```
<!ENTITY  %   ISOlat1

  PUBLIC   "ISO 8879:1986//ENTITIES Added Latin 1//EN" >
```

Respecto al tipo de entidad externa, puede ser (sólo para entidades generales):

- **SUBDOC:** Es texto SGML, pero tiene su propia DTD.
- **NDATA:** Son gráficos u otro tipo de datos no codificados siguiendo las normas de SGML.

- **CDATA:** Son datos de tipo carácter.
- **SDATA:** Son datos con caracteres específicos de un sistema.

Ejemplo `<!ENTITY entidad1 SYSTEM "entidad1.sgm" SUBDOC>`

Si los datos del archivo no han sido codificados usando SGML se debe especificar, después del tipo, un nombre de notación (puede ser especificada para entidades de tipo `NDATA`, `CDATA` y `SDATA`). El nombre de notación debe haber sido declarado en la misma DTD.

Ejemplo `<!ENTITY figura SYSTEM "figura.gif" NDATA GIF>`

Las Figuras 2.10 y 2.11 resumen las posibilidades que hay a la hora de declarar una entidad general o una entidad paramétrica.

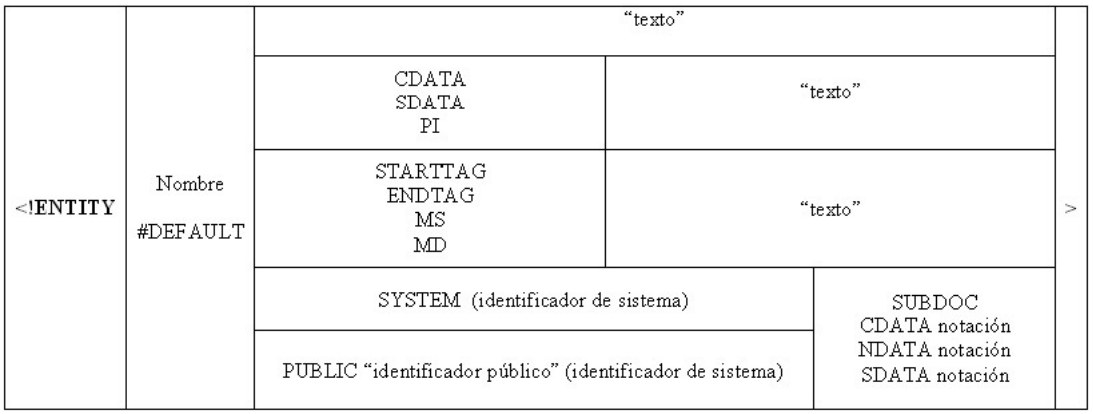

Figura 2.10: Declaración de entidad general.

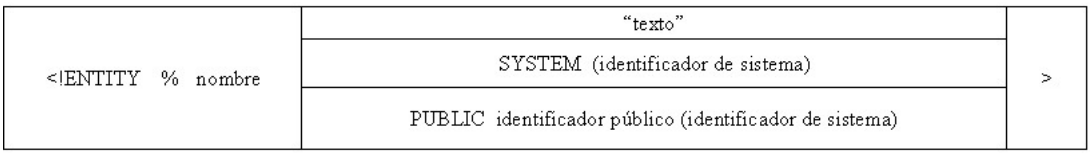

Figura 2.11: Declaración de entidad paramétrica.

2.2.3.4. Declaración de notación

Muchas veces es necesario incluir en un documento SGML datos que han sido creados de otro modo. Para cada uno de estos datos debe indicarse el modo correcto de procesarlos. Con este fin se usa la declaración de notación. Se verá brevemente la forma de hacerlo. La sintaxis que debe usarse en la declaración de una notación es la siguiente:

```
<!NOTATION nombre SYSTEM/PUBLIC ''identificador_externo''>
```

Se comentará ahora, el significado de cada término.

MDO	Delimitador con el que comienza una declaración ("<!").
NOTATION	Palabra reservada que distingue la declaración de una notación.
nombre	Nombre de la notación.
SYSTEM/PUBLIC	Tienen el mismo significado visto para las entidades.
identificador_externo	Para las notaciones declaradas como `SYSTEM` será la ruta para llegar hasta el archivo, para las declaradas como `PUBLIC` será un texto normalizado que aporta información sobre la entidad.
MDC	Delimitador de fin de declaración(">").

Ejemplo:
```
<!NOTATION postscript SYSTEM "eps.bat" >
<!ENTITY figura SYSTEM "figura.eps" NDATA postscript >
```

Cuando se usa un nombre de notación en una declaración de entidad, también pueden añadirse atributos que serán pasados al sistema como parámetros. Esos atributos se definen con una declaración ATTLIST similar a la de los atributos de un elemento, pero el nombre de notación irá precedido por la palabra #NOTATION. Por ejemplo:

```
<ATTLIST #NOTATION postscript width NUTOKEN #IMPLIED>
```

La declaración de entidad con atributos sería como la siguiente:

```
<!ENTITY  figura SYSTEM "figura.eps" NDATA postscript [width=''5'']>
```

2.2.4. Otros aspectos del lenguaje

2.2.4.1. Caracteres y referencias de carácter

En un documento SGML, puede emplearse cualquier conjunto de caracteres, siempre que sea suficientemente extenso como para poder representar los caracteres que se usan en el marcado y en los datos de usuario. Ya se ha visto que el conjunto de caracteres usados se especifica en el apartado CHARSET de la declaración SGML. Por defecto se usa el conocido como ISO 646.

No todos los caracteres del conjunto podrán emplearse en cualquier parte del documento. Habrá que diferenciar entre caracteres delimitadores, caracteres para los identificadores, caracteres con los que puede comenzar un identificador, caracteres de control y caracteres de datos.

- Los caracteres delimitadores son los que diferenciarán los caracteres de marcado de los caracteres que forman los datos de usuario. Se asignan en la sintaxis concreta.
- Los caracteres que pueden usarse en los identificadores de elemento, atributo o sección se conocen como "name characters". Siempre forman parte de este conjunto las letras (de la A a la Z, mayúsculas o minúsculas) y los dígitos (del 0 al 9). La sintaxis concreta de referencia añade también el punto "." y el guión "-".

- Los caracteres con los que puede empezar un identificador forman parte de un conjunto más limitado que el anterior llamado "name start character". Básicamente son las letras. Pueden añadirse otros caracteres según la variante de sintaxis concreta que se use.
- Pueden usarse los caracteres de control conocidos como RS, RE y el espacio. La sintaxis concreta de referencia añade también el tabulador.
- Los caracteres que pueden usarse como datos de usuario se conocen como "data characters". Aquí pueden añadirse símbolos matemáticos y de puntuación que no formaban parte de los "name characters".

En SGML, los caracteres delimitadores no son caracteres fijos, sino que pueden ser asignados en la sintaxis concreta que se defina según las necesidades del usuario. Sin embargo, a veces es necesario utilizar alguno de los caracteres delimitadores como parte de los datos. Para poder hacerlo, y que estos caracteres sean considerados datos y no delimitadores, deben usarse referencias. Si, por ejemplo, se definen las siguientes entidades:

```
<!ENTITY   amp   ''&''  >

<!ENTITY   lt   ''<''  >

<!ENTITY   gt   ''>''  >
```

Cada vez que alguno de esos caracteres deba aparecer en los datos de usuario, se insertarán mediante sus referencias : `&,&lt,">" y "&quot;";`

También será necesario usar referencias de carácter cuando haya que introducir caracteres que no existan en el teclado o caracteres que no sean SGML y no puedan ser introducidos directamente. Para añadir un carácter de esta manera, se usa su código decimal. La referencia estará formada por los caracteres `'&#'`, seguidos del código decimal y terminada en `';'`. Por ejemplo, las dos declaraciones siguientes son equivalentes:

```
<!ENTITY   amp   ''&''  >

<!ENTITY   amp   ''&#38;''  >
```

Existen referencias de carácter con nombres predefinidos para los caracteres de control como las siguientes: `&#RS;`, `&#RE;` y `&#SPACE;`. Puede definirse más nombres para referencias si se añaden en la sintaxis concreta.

2.2.4.2. Comentarios

Hay un tipo de declaración que se usa para insertar comentarios, y que puede aparecer tanto en la DTD como en la instancia de documento. Debe tener la forma siguiente:

```
<!-- comentario -->
```

Donde se ven los términos que se comentan a continuación:

MDO	Delimitador con el que comienza una declaración (”<!”).
COM	Caracteres que marcan el inicio o el fin de un comentario (”- -”).
Comentario	Una cadena de caracteres.
MDC	Delimitador de fin de declaración (“>”).

No deben dejarse espacios en blanco entre los delimitadores MDO y COM ni entre COM y MDC. Es decir, el siguiente ejemplo no sería válido:

```
<! -- comentario -- >
```

Los comentarios pueden aparecer dentro de otra declaración (aunque nunca dentro de un grupo que esté entre paréntesis ni un grupo de modelo de contenido). En este caso, solo deben ir precedidos y seguidos por los delimitadores COM. Por ejemplo:

```
<!ELEMENT (p|titulo)          - o   (#PCDATA) --comentario-->
```

Si es necesario, también pueden incluirse varios comentarios en una misma declaración de comentario.

Finalmente, si sólo se desea incluir una línea en blanco, puede usarse una forma reducida de la declaración de comentario que consiste en un MDO inmediatamente seguido por un MDC (`<!>`).

2.2.4.3. Instrucciones de proceso

Las instrucciones de proceso son instrucciones para el sistema que procesa el documento, estando escritas en el lenguaje que éste necesite. Desde el punto de vista de SGML sólo es necesario saber diferenciarlas del resto del marcado, lo que se consigue con los delimitadores PIO y PIC , que en la sintaxis concreta de referencia son los caracteres ’’<?’’ y ’’>’’ .

```
<? instrucción >
```

2.2.4.4. Secciones marcadas

Una sección marcada es una parte del documento que necesita ser procesada de una forma especial, por ejemplo, ser ignorada en ciertos procesos o deshabilitar en ella la búsqueda de caracteres de marcado. La forma de declarar una sección marcada es la siguiente:

```
<![ tipo [ sección marcada ]]>
```

Está formada por los siguientes componentes:

MDO	Delimitador con el que comienza una declaración (”<!”).
DSO	Caracteres que marcan el inicio de un subconjunto de declaración (”`[`”).
`tipo`	Tipo de sección marcada.
DSO	El segundo DSO indica el inicio de la sección marcada (”`[`”).
MSC	Delimitador de fin de sección marcada (”`]]`”).
MDC	Delimitador de fin de declaración (“>”).

Debe comentarse que es necesario tener algún tipo de carácter separador entre el primer DSO y el tipo de sección, sin embargo, no puede haber ningún carácter separador en la cadena formada por los separadores: ''`<![` '' ni en la cadena: ''`]]>`''.

Los tipos de sección marcada que se pueden tener son:

- `IGNORE` : Indica que el contenido debe ser ignorado por el parser durante el análisis. Eso no significa que pueda contener caracteres no válidos en SGML.
- `INCLUDE`: Indica que el contenido sí debe ser analizado. Es el tipo de sección marcada que se usa por defecto si no se especifica ninguno.
- `CDATA`: Indica que el contenido debe ser tratado como datos de usuario (no deben buscarse en él caracteres de marcado).
- `RCDATA`: Indica que el contenido debe ser tratado como datos, pero sí deben reemplazarse las referencias a caracteres o a entidades.
- `TEMP`: Indica que el contenido es parte del documento de forma temporal.

Pueden usarse varios de estos tipos al principio de una misma sección marcada. En ese caso, la prioridad más alta sería para `IGNORE` seguida de `CDATA`, `RCDATA` e `INCLUDE`, en ese orden.

Los dos primeros tipos (`IGNORE` e `INCLUDE`), suelen usarse cuando un documento va a ser procesado por dos sistemas distintos y requieren instrucciones distintas, o bien cuando quieren generarse dos versiones distintas de un documento sin tener que duplicar la parte común. Sólo hay que añadir la parte del documento que difiere en una sección marcada. Por ejemplo [19]:

```
<![ IGNORE [<?instrucción para el sistema B>]]>

<![ INCLUDE [<?instrucción para el sistema A>]]>
```

Los tipos `CDATA` y `RCDATA` se usan cuando el usuario necesita usar etiquetas o caracteres propios del marcado en sus datos iniciales. Si necesita que las entidades y las referencias de caracteres sean resueltas, usará `RCDATA`, y si no, `CDATA`. En el siguiente ejemplo, las etiquetas del título no serán tomadas en cuenta como parte del marcado.

```
<![ CDATA [<titulo>SGML</titulo>]]>
```

El tipo `TEMP` puede usarse para marcar una sección que es temporal, de forma que sea fácil identificarla y eliminarla si es necesario.

```
<seccion>

<titulo>Minimización</titulo>

<p>Un documento...<![ TEMP [sin terminar]]></p>

</seccion>
```

2.3. XML

En esta sección se estudia XML, el lenguaje extensible de marcas, un lenguaje de marcado que surge como subconjunto de SGML en un intento de simplificarlo. En la introducción al lenguaje se incluirán los objetivos con los que se diseñó el lenguaje, se repasarán los conceptos de elemento, DTD y atributo, que ya se vieron en la Sección 2.2, y la diferencia entre documentos válidos y documentos bien formados. A continuación se describirá el documento XML, las partes que lo componen y la sintaxis de cada una, la DTD, haciendo hincapié en las restricciones de una DTD para XML respecto a las de SGML y los posibles inconvenientes que ello genera, y finalmente, otros aspectos del lenguaje que sufren modificaciones respecto a SGML, entre ellos el uso de las características adicionales.

2.3.1. Introducción al lenguaje XML

XML es un metalenguaje con el que pueden definirse lenguajes de marcado personalizados que se adapten a las necesidades de distintos tipos de documentos. Fue desarrollado por el W3C en 1996 (y publicado en 1998) como una versión reducida de SGML.

La necesidad de un nuevo lenguaje surge con el avance de la Web. En los comienzos de la *World Wide Web* se popularizó el lenguaje de marcado HTML, sin embargo, con HTML el conjunto de etiquetas que pueden ser usadas está predefinido, lo que supone una gran limitación a pesar de que el conjunto de etiquetas se esté ampliando constantemente. La alternativa era usar SGML, con el que pueden describirse estructuras más ricas, pero se consideró que era demasiado complejo para los usuarios, para implementar programas que lo procesen y para el tipo de información que suele usarse en la Web, por ello se optó por una simplificación de SGML. XML tiene buena parte de la potencia de SGML eliminando en gran medida su complejidad, sin embargo, en su desarrollo se eliminaron algunas de las características que daban flexibilidad a SGML, aquellas que pudiesen producir ambigüedades o dificultar el tratamiento de los documentos, por eso no puede decirse que XML sustituya a SGML.

Las diferencias más notables entre ambos lenguajes se aprecian en el desarrollo de las DTDs, de hecho puede ser imposible encontrar la forma de expresar algunas de las restricciones de una DTD que con SGML serían sencillas, y en la eliminación de las características opcionales de SGML, entre ellas las minimizaciones.

El objetivo fundamental de XML, por tanto, es permitir que SGML pueda ser servido, recibido y procesado en la web en la misma manera que es posible con HTML[41]. Oficialmente, los objetivos que se plantearon en el diseño de XML fueron los siguientes [41]:

- XML debe poder usarse en Internet.
- XML debe soportar una amplia variedad de aplicaciones.
- XML debe ser compatible con SGML.
- Los programas para procesar documentos XML deben ser sencillos.
- El número de características adicionales en XML debe el mínimo posible.

- Los documentos XML deben poder ser leídos directamente por un usuario.
- El diseño de XML debe ser rápido, formal y conciso.
- Los documentos XML deben ser fáciles de crear.
- La brevedad en la marcación no se considerará importante.

El intercambio de información estructurada no es la única utilidad de XML, además puede usarse, por ejemplo, como formato de almacenamiento, es el caso de las bases de datos nativas XML.

Al ser XML un subconjunto de SGML, los conceptos fundamentales son los mismos. En la sección de introducción a SGML (2.2.1) ya se definieron los conceptos de elemento, atributo y DTD. Se repasarán aquí estos conceptos básicos y será en la siguiente sección cuando se vean las diferencias de sintaxis existentes entre SGML y XML.

En un documento marcado, la información está separada en elementos y estos vienen delimitados por etiquetas o marcas, por tanto, un elemento es cada una de las partes en que se divide la información junto con sus etiquetas. Por ejemplo, uno de los elementos de un documento puede ser su título y podría tener la siguiente forma:

```
<titulo>XML</titulo>
```

El elemento está formado por una etiqueta de inicio (`<titulo>`), el contenido (el propio título, `XML`) y una etiqueta de fin (`</titulo>`).

Una manera de poder aportar información adicional sobre un elemento es usar atributos. Si, por ejemplo, otro de los elementos del documento fuese el elemento capítulo, y se quisiera especificar el estado de desarrollo en que se encuentra, bastaría con incorporar un atributo en la etiqueta de inicio del elemento. Se haría de la siguiente manera:

```
<capitulo estado=''borrador''>
```

Un elemento puede contener a otros formando una estructura jerárquica. Esa estructura puede venir definida por una DTD, aunque no siempre será así, puesto que en XML la DTD es opcional. Si el documento tiene una DTD y sigue su estructura, es llamado documento válido, si carece de DTD, debe respetar igualmente la jerarquía en las etiquetas, todos los elementos deben estar correctamente anidados. En ese caso se dice que el documento está bien formado.

Además de la DTD, que proviene de SGML, actualmente hay otras formas de especificar la estructura del documento mediante otro tipo de esquemas, por ejemplo, los XML Schemas, que usan la sintaxis de XML. Si el documento cumple las reglas de alguno de estos esquemas también se considera documento válido.

2.3.2. Documento XML

Un documento XML puede tener un prólogo y debe tener una instancia de documento. En este apartado se describirán ambos componentes. Por un lado la sintaxis del prólogo, que puede contener la declaración XML y la declaración de tipo de documento (y por

tanto, la DTD), y por otro, la instancia del documento, que es la parte que contiene los datos originales.

Es interesante comentar que los documentos SGML empiezan con una declaración SGML, la cual permite configurar a medida algunas características del documento, por ejemplo, el conjunto de caracteres que se usarán o las palabras clave. En XML, la declaración SGML está prefijada y no debe aparecer explícitamente, con lo que se pierden todas esas opciones de personalización en busca de mayor sencillez.

2.3.2.1. Prólogo

El prólogo debe ser lo primero en aparecer en el documento en el caso de que lo incluya. Si el documento contiene un prólogo, debe aparecer antes del primer elemento. Está formado por la declaración XML y/o la declaración de tipo de documento.

- **Declaración XML**

La declaración XML, aunque opcional, es muy recomendable. Si se incluye, debe especificar la versión de XML usada. Según la especificación, debe seguir la sintaxis que se indica a continuación:

```
<?xml version encoding standalone?>
```

<?: Delimitador con el que comienza una declaración.

`xml`**:** distingue la declaración XML.

`version`**:** Debe tener como valor la versión de XML usada.

`encoding`**:** Codificación del documento.

`standalone`: Indica si el documento depende o no de declaraciones externas (mediante los valores "no" o "yes" respectivamente).

?>: Delimitador de fin de declaración.

Un ejemplo de lo anterior sería la siguiente declaración:

```
<?xml version=" 1.0 " encoding=" UTF-8 " standalone= " yes "?>
```

Tanto la declaración de codifiación (`encoding`)como la de independiente (`standalone`) son opcionales, además el valor por defecto para la codificación es UTF-8, por lo que en este caso no hubiese sido necesaria. Al declararlo como independiente, el analizador debe indicar un error si se hace referencia, por ejemplo, a una DTD externa.

- **Declaración de tipo de documento**

La declaración de tipo de documento es la sección que contiene la definición de tipo de documento, o bien, nos dice dónde podemos encontrarla. Por su parte, La definición de tipo de documento (DTD) contiene las normas que deben cumplir las instancias de un tipo de documento concreto: los elementos que pueden usarse, sus posibles tipos de contenido, sus atributos, las entidades, etc. Se verá su sintaxis en el apartado 2.3.3.

Si se quiere incluir la DTD en el mismo archivo que la declaración, se usa un subconjunto de la declaración, en caso contrario, se usa un identificador de un archivo almacenado en el sistema o un identificador público. También pueden usarse ambas posibilidades combinadas para ampliar la DTD de un archivo con nuevos elementos definidos en el subconjunto de la declaración. Para diferenciarlos se habla de subconjunto interno y subconjunto externo de la declaración.

Su sintaxis es la misma que en SGML, se incluye en la Sección 2.2.2.2, pero hay que señalar dos restricciones importantes que afectan al subconjunto interno de la declaración:

- Si se usan referencias a entidades paramétricas, éstas deben reemplazar declaraciones completas y no a partes de las mismas.
- No pueden usarse secciones marcadas.

2.3.2.2. Instancia de documento

La instancia de documento contiene los datos originales junto con las etiquetas usadas para marcarlos. En XML el documento puede reducirse sólo a eso, de hecho, si no aparece una DTD, pueden usarse las etiquetas que se necesiten sin definirlas. Tanto si el documento incluye una DTD como si no lo hace, la estructura debe respetarse estrictamente a la hora de etiquetar los elementos, es decir, debe haber un elemento raíz, cada elemento debe tener su etiqueta de inicio, su contenido, y su etiqueta de fin (excepto que sea elemento vacío), y todos deben estar correctamente anidados. Es la forma de que la aplicación que procese el documento entienda la estructura correctamente aunque no se incluya una DTD.

Las etiquetas siguen básicamente la misma sintaxis que en SGML. Los caracteres separadores son los mismos que define la sintaxis concreta de referencia de SGML, aunque en SGML podía variarse la sintaxis concreta y usar otros caracteres STAGO, TAGC y ETAGO. Aquí se pierde esa posibilidad. La sintaxis sería la siguiente:

- La etiqueta de inicio debe comenzar con el carácter '<' seguido del identificador de elemento y terminar con el carácter '>'. En el caso de que el elemento tenga atributos, estos deben ir antes del carácter de cierre. Por ejemplo: `<informe estado= ‘‘borrador’’>`. Los atributos sólo pueden especificarse una vez y en cualquier orden.
- La etiqueta de fin es similiar a la etiqueta de inicio salvo porque debe comenzar con los caracteres '</' en vez de '<' y porque no lleva atributos. Por ejemplo: `</informe>`.

Los elementos vacíos tienen una sintaxis distinta a la de SGML. En SGML un elemento vacío no puede tener etiqueta de fin, aquí puede usarse una etiqueta de inicio inmediatamente seguida por una etiqueta de fin, o bien, utilizar una única etiqueta, la etiqueta de elemento vacío, que tiene la misma forma vista para la etiqueta de inicio, pero acaba con los caracteres '/>' en vez del carácter '>'. Los dos elementos siguientes serían equivalentes:

```
<imagen></imagen>
```

```
<imagen/>
```

Como parte del contenido de un elemento, según la recomendación [41], pueden aparecer: otros elementos, datos de usuario, comentarios, instrucciones de proceso, referencias o secciones CDATA.

Si se incluye una DTD debe respetarse el tipo de contenido que se declare para cada elemento, sus atributos y las entidades que pueden referenciarse, si no se incluye, se usarán los elementos y atributos que se necesiten y podrán usarse las entidades predefinidas que se ven en la Sección 2.3.4.1. Si el documento es válido, debe declarar todas las entidades que utiliza, incluso las predefinidas.

2.3.3. Implementación de la DTD

En esta sección se verán las diferencias entre XML y SGML a la hora de declarar elementos, atributos, entidades y notaciones. En la sección dedicada a SGML se trabajó con la DTD que se repite en la Figura 2.12, para poder usarla con documentos XML habría que modificarla ligeramente. La DTD modificada se muestra en la Figura 2.13. La explicación de estas modificaciones se ve a lo largo de la sección.

```
<!ELEMENT informe  - -   (titulo,seccion+,bibliografia)>
<!ELEMENT seccion  - o   (titulo,p+)>
<!ATTLIST seccion estado (borrador|completo) borrador>
<!ELEMENT bibliografia  - o   (p+)>
<!ELEMENT (p|titulo) - o   (#PCDATA)>
```

Figura 2.12: Ejemplo de DTD (válida en SGML) para un informe.

```
<!ELEMENT informe  (titulo,seccion+,bibliografia)>
<!ELEMENT seccion  (titulo,p+)>
<!ATTLIST seccion estado (borrador|completo) 'borrador'>
<!ELEMENT bibliografia  (p+)>
<!ELEMENT p   (#PCDATA)>
<!ELEMENT titulo  (#PCDATA)>
```

Figura 2.13: Ejemplo de DTD (válida en XML) para un informe.

2.3.3.1. Declaración de elemento

En la declaración de cada elemento se describen su nombre y todas las posibilidades que se tienen en cuanto a su contenido. La sintaxis básica de la declaración es la misma que

en SGML, salvo porque no aparecen los dos caracteres que indicaban la posibilidad de omitir las etiquetas de inicio o de fin del elemento, en XML no se tiene dicha posibilidad. La sintaxis sería la siguiente:

```
<!ELEMENT       nombre      contenido>
```

En SGML se tenía la posibilidad de declarar varios elementos a la vez, aquí es necesario declararlos de forma independiente aunque coincidan sus modelos de contenido. Si en la DTD para documentos SGML se utilizó el siguiente ejemplo:

```
<!ELEMENT (p|titulo)          - o   (#PCDATA)>
```

El mismo ejemplo ahora quedaría de la siguiente forma:

```
<!ELEMENT p       (#PCDATA)>
<!ELEMENT titulo  (#PCDATA)>
```

Al especificar el tipo de contenido las posibilidades se reducen:

- No pueden usarse las palabras reservadas CDATA y RCDATA como contenido del elemento. Habría que usar #PCDATA en su lugar, aunque ello supone que no podrán usarse los caracteres de marcado en el contenido, a no ser que se haga con secuencias de escape o con una sección CDATA.
- Cuando se usen grupos no pueden incluir el conector '&', con lo que no pueden especificarse grupos donde los elementos puedan aparecer en cualquier orden.
- No pueden utilizarse grupos de inclusión o exclusión.
- En los modelos de contenido mixto no puede limitarse el orden ni el número de veces que aparecen los elementos. Deben usarse grupos opcionales y repetibles, separados por el conector "|", y en los que #PCDATA aparezca en primer lugar, por ejemplo: `(#PCDATA|elemento1|elemento2)*`, permitiéndose el número de elementos que se desee.

2.3.3.2. Declaración de la lista de atributos

Mediante la declaración de la lista de atributos es posible definir los atributos que puede tener cada elemento, los tipos de valores que pueden tener los atributos y sus valores por defecto.

La siguiente sería una declaración de atributos válida en XML.

```
<!ATTLIST informe
          estado (borrador|completo) 'borrador'>
```

Se declara un atributo `estado`, que puede tomar los valores `borrador` o `completo`, y tomará el valor `borrador` si no se especifica ninguno. Como se ve en el ejemplo, en XML es necesario delimitar el valor por defecto de un atributo con comillas (dobles o simples), además el único conector permitido para especificar el grupo de posibles valores es el conector OR ('|'). Tampoco se puede usar una misma declaración ATTLIST para especificar los atributos de varios elementos a la vez mediante un grupo.

No son las únicas restricciones que impone XML, también están restringidas las palabras reservadas que se vieron para especificar los tipos de atributos y sus valores por defecto:

- Para especificar los tipos de atributos no pueden usarse las siguientes palabras que sí son válidas en SGML:
  - NAME o NAMES.
  - NUMBER o NUMBERS.
  - NUTOKEN o NUTOKENS.
- Para especificar los valores por defecto de un atributo no pueden usarse las siguientes palabras reservadas:
  - CURRENT
  - CONREF

2.3.3.3. Declaración de entidad

Una entidad es una unidad de almacenamiento que puede contener un conjunto de caracteres, un trozo de la DTD o un archivo externo [41]. Cada documento XML tiene una entidad documento, y en ella pueden hacerse referencias a otras entidades, lo que causará que el procesador las incluya como parte del documento. El contenido de las entidades puede ser procesado, en ese caso se consideraría una parte más del documento XML, o sin procesar, en ese caso su contenido posiblemente no sería texto o bien sería texto que no fuese XML.

Tal y como se vio en la sección dedicada a SGML, las entidades pueden clasificarse como entidades generales o paramétricas, y éstas a su vez, como entidades internas o externas. El siguiente sería un ejemplo de entidad interna y general.

```
<!ENTITY  entidad1 "texto que se reemplaza">
```

La entidad con nombre `entidad1` es una entidad interna porque su contenido (`"texto que se reemplaza"`) está incluído en su declaración, y es general porque se usará (se hará referencia a ella causando que su contenido sea incluido en el punto donde se referencie) dentro del contenido del documento. En SGML se vieron distintos tipos de entidades generales, en XML no pueden usarse entidades de tipo CDATA, SDATA ni PI.

Las entidades que contienen parte de una declaración son llamadas entidades paramétricas. Éstas deben ser referenciadas dentro de una declaración. La entidad que se declara en el siguiente ejemplo es una entidad interna y paramétrica.

```
<!ENTITY  % entidad2   "#PCDATA | seccion | bibliografia">
```

El contenido de `entidad2` también está incluido en su declaración. Si el contenido de una entidad está almacenado en un fichero externo, la entidad sería una entidad externa, como la del siguiente ejemplo:

```
<!ENTITY  entidad3      SYSTEM     "c:\SGML\ejemplos\entidad3.sgml" >
```

En la declaración de `entidad3` se incluye la palabra reservada `SYSTEM`, que indica que el contenido de la entidad está almacenado en el mismo sistema en que se crea el documento, seguida de una cadena de caracteres que contiene la ruta para llegar hasta el fichero. Si en lugar de `SYSTEM` apareciese la palabra reservada `PUBLIC` iría seguida de un identificador público formal y después, también debería incluir el identificador de sistema con la URI para acceder a la entidad (en SGML era suficiente con el identificador público). Por ejemplo:

```
<!ENTITY% ISOlat1 PUBLIC "ISO 8879:1986//ENTITIES Added
  Latin 1//EN//XML" "ISOlat1.sgml" >
```

Hasta ahora todos los ejemplos vistos han sido ejemplos de entidades procesadas, si se tiene una entidad externa con un contenido no procesado (las entidades internas son siempre procesadas), debe incluir la palabra reservada NDATA seguida del nombre de una notación previamente declarada. Por ejemplo:

```
<!ENTITY  figura      SYSTEM     "figura.gif"   NDATA   GIF>
```

En SGML se vieron otros tipos de entidad externa (SDATA, CDATA, SUBDOC), aquí ninguno de ellos es válido.

La sintaxis de la declaración de entidad fue comentada más extensamente en la sección dedicada a SGML, aquí sigue siendo válida, salvo por las excepciones comentadas.

La forma de referenciar una entidad en XML también presenta alguna restricción respecto a la de SGML. A continuación se muestran ejemplos de referencias a una entidad general y una paramétrica:

- Referencia a una entidad general: **&entidad1;**
- Referencia a una entidad paramétrica: **%entidad2;**

En SGML existía la posibilidad de usar como fin de referencia un carácter fin de línea u otro carácter (como un espacio), que no fuese parte de un nombre válido de entidad, aquí no es posible. El carácter REFC (';') debe aparecer siempre como fin de una referencia. Por otro lado, los lugares en que puede referenciarse una entidad según sea paramétrica o general, interna o externa, procesada o no procesada, están limitados en XML. En la recomendación [41], se definen los contextos en los que puede aparecer cada tipo de referencia y el comportamiento que debe tener el procesador XML en cada caso. Por ejemplo, no pueden aparecer referencias externas en el valor de un atributo.

2.3.3.4. Declaración de notación

Las notaciones identifican los distintos tipos de formato que se usan en el documento XML (como el formato de entidades no procesadas) y que no son XML. La declaración de notación incluye la localización de la aplicación que es capaz de procesar ese tipo de datos. La sintaxis de la declaración coincide con la misma en SGML, pero en XML no pueden añadirse atributos a una notación. Por ejemplo:

```
<!NOTATION postscript SYSTEM "eps.bat" >
```

2.3.4. Otros aspectos del lenguaje

2.3.4.1. Caracteres y referencias de carácter

En XML el conjunto de caracteres legales está formado por los caracteres Unicode e ISO 10646, pero no todos los caracteres de dichos estándares están permitidos. Son caracteres legales los espacios en blanco, el tabulador, el retorno de carro, el avance de línea y los caracteres gráficos. La mayoría de los caracteres no gráficos están prohibidos.

Por otro lado, que un carácter sea legal en XML no quiere decir que pueda usarse en cualquier parte del documento. Hay caracteres que pueden usarse como parte del marcado pero no como parte de los datos, otros que no pueden usarse en los identificadores e incluso algunos que pueden formar parte de un identificador, pero no ser el primer carácter de uno de ellos. Para ver cada caso se empieza diferenciando entre los caracteres que pueden considerarse caracteres de marcado y los que son caracteres de datos:

- Los caracteres de marcado son aquellos que forman parte de etiquetas de comienzo, de fin, de elementos vacíos, referencias de entidad, referencias de carácter, comentarios, delimitadores de secciones CDATA, declaraciones de tipo de documento, instrucciones de procesamiento, declaraciones XML y espacios en blanco que existan fuera de la entidad del documento. Los caracteres delimitadores son los que diferenciarán los caracteres de marcado de los caracteres que forman los datos de usuario.
- Caracteres de datos son todos aquellos que no forman parte de los caracteres de marcado. Si se habla del contenido de un elemento, será cualquier cadena de caracteres que no contenga el delimitador de comienzo de ningún tipo de marcado, ni el delimimitador de cierre de una sección CDATA, y si se habla del interior de una sección CDATA, será cualquier carácter excepto el delimitador de cierre de sección CDATA[41].

Hecha esta distinción, puede decirse que los caracteres '<' y '&' no pueden ser usados como datos en un documento (salvo si se usa una referencia de carácter, que se comenta a continuación, o si están dentro de una sección CDATA), pues podrían confundirse con el inicio de datos de marcado. El carácter '>' tampoco debe usarse como tal (debe usarse una referencia de carácter) después de una cadena ']]' si no quiere confundirse con el final de una sección CDATA.

Las referencias de caracteres ya se vieron en la sección dedicada a SGML. Son útiles para añadir al documento caracteres que no puedan ser introducidos directamente, por ejemplo, porque no existan en el teclado, o para usar como parte de los datos un carácter que pueda ser confundido con un carácter de marcado, como los carácteres '<' y '&'. También deben usarse referencias de carácter para añadir las comillas dobles en el caso de que se hayan usado para delimitar el valor de un atributo, si se han usado comillas simples para delimitarlo no será necesario (ocurre lo mismo en el caso opuesto, es decir,

si se han usado comillas dobles para delimitar el valor, las comillas simples pueden usarse en el contenido sin necesidad de referenciarlas).

Las referencias de caracteres pueden construirse de dos maneras: bien con los caracteres '&#', seguidos del código decimal del carácter que se quiere introducir y, terminada en ';' o bien con los caracteres '&#x', seguidos del código hexadecimal y también terminada en ';'.

En XML hay cinco entidades que ya están predefinidas para representar este tipo de caracteres. Estas entidades serán reconocidas incluso si no están declaradas, pero un documento XML válido debería declararlas igualmente. Pueden verse en la Figura 2.14.

Carácter	**Referencia**
&	&amp;
<	<
>	>
'	&apos;
"	&quot;

Figura 2.14: Referencias de carácter en XML.

La forma de declararlas en un documento sería la siguiente:

```
<!ENTITY lt "&#38;#60;">

<!ENTITY gt "&#62;">

<!ENTITY amp "&#38;#38;">

<!ENTITY apos "&#39;">

<!ENTITY quot "&#34;">
```

Como se ha comentado anteriormente, los caracteres que forman parte de un identificador de elemento, atributo o sección también están limitados. Un identificador puede contener básicamente: letras (de la A a la Z, mayúsculas o minúsculas), dígitos (del 0 al 9), guiones, subrayados(underscores) y puntos. El carácter de dos puntos está reservado para trabajar con espacios de nombres. Los caracteres con los que puede empezar un identificador están aún más limitados, básicamente, deben comenzar por letras o guiones bajos. Los nombres que comienzan con la cadena "xml" o las combinaciones de estas letras en mayúsculas o minúsculas están reservados. El conjunto de caracteres completo que pueden usarse para comenzar un identificador o para formar parte del mismo se incluye en la Figura 2.15. Es importante señalar también que, a diferencia de lo que ocurre en SGML, en XML se debe diferenciar entre mayúsculas y minúsculas al hablar de identificadores de elementos y atributos.

```
NameStartChar ::= ":" | [A-Z] | "_" | [a-z] | [#xC0-#xD6] | [#xD8-#xF6] |
[#xF8-#x2FF] |
                  [#x370-#x37D] | [#x37F-#x1FFF] | [#x200C-#x200D] |
[#x2070-#x218F] |
                  [#x2C00-#x2FEF] | [#x3001-#xD7FF] | [#xF900-#xFDCF] |
[#xFDF0-#xFFFD] |
                  [#x10000-#xEFFFF]
NameChar ::= NameStartChar | "-" | "." | [0-9] | #xB7 | [#x0300-#x036F] |
[#x203F-#x2040]
```

Figura 2.15: Caracteres permitidos en identificadores.

2.3.4.2. Comentarios

Los comentarios en XML tienen la misma forma que en SGML, aunque con las siguientes restricciones:

- No puede incluirse un comentario dentro de otra declaración.
- No puede usarse la forma reducida vista en SGML para comentarios vacíos (<!>).
- No puede haber varios comentarios dentro de una misma declaración de comentario.

Por ejemplo:

```
<!-- Esto sería un comentario válido -->
```

```
<!ELEMENT x - - (A,B) – Esto sería un comentario no válido –>
```

2.3.4.3. Instrucciones de proceso

Las instrucciones de proceso son instrucciones para las aplicaciones que procesan el documento. Para delimitarlas se usan los delimitadores PIO (''<?'') y PIC(''?>'') (El delimitador PIC no coincide con el que se usa en SGML siguiendo la sintaxis concreta de referencia). Deben comenzar con un objetivo, el PITarget, que identifique la aplicación a la que se dirige. El objetivo no debe ser xml ni XML (ni cualquiera de las combinaciones de estos caracteres variando mayúsculas y minúsculas), salvo que sea la propia declaración XML.

```
<? PITarget instrucción ?>
```

2.3.4.4. Secciones CDATA

Las secciones CDATA son fragmentos del documento en los que el analizador no buscará caracteres de marcado (excepto la cadena de caracteres que indica el fin de sección '']]>''). Se usan para delimitar información en la que existen caracteres de marcado, pero no se desea que se interpreten como tales, sino como información de usuario.

Estas secciones se vieron en la sección dedicada a SGML como uno de los tipos de secciones marcadas. El resto de tipos de sección marcada que se vieron allí no pueden ser usados en XML. Por esa razón, no pueden usarse varios tipos al principio de una sección marcada. Tampoco puede usarse una sección marcada en la que no se indique el tipo de sección, ni pueden ser anidadas.

La sintaxis sigue siendo la misma vista en SGML siempre que el tipo de sección se sustituya por CDATA.

```
<![ CDATA [ sección marcada ]]>
```

El ejemplo visto para SGML seguiría siendo válido aquí:

```
<![ CDATA [<titulo>SGML</titulo>]]>
```

En este ejemplo, las cadenas ''`<titulo>`'' y ''`</titulo>`'' no son consideradas etiquetas sino datos de carácter.

2.3.4.5. Características adicionales

Uno de los objetivos de diseño de XML fue reducir al mínimo las características adicionales descritas para SGML, entre ellas las características de minimización que se describen en el Capítulo 3. En su diseño se consideró que las limitaciones de memoria no eran ya tan importantes como la complejidad que introducen estas características, y por ello se prohíbe el uso de DATATAG, OMITTAG, RANK, SHORTREF y USEMAP. Tampoco se consideran esenciales, y por tanto se eliminan, las características LINK, CONCUR, SUBDOC y FORMAL. La característica SHORTTAG se mantiene con el objetivo de permitir el uso de la etiqueta de elemento vacío que no existe como tal en SGML, para eso se hace uso de la opción NET (Null End Tag), pero ninguna de las otras posibilidades de SHORTTAG está permitida.

En SGML, la opción NET, permite sustituir la etiqueta de fin de elemento por un solo carácter, el separador "null end tag", que suele ser "/", siempre que el separador final de la etiqueta de inicio de dicho elemento se sustituyese por ese mismo separador. Por ejemplo, un elemento título, que podría ser el siguiente:

```
<titulo>SGML</titulo>
```

Puede ser reducido de la siguiente manera:

```
<titulo/SGML/
```

En XML aquellos elementos que carezcan de contenido pueden expresarse con una única etiqueta que tiene una sintaxis especial: tiene la misma forma vista para la etiqueta de inicio, pero acaba con los caracteres '/>' en vez del carácter '>'. Para poder permitirlo se aprovecha la opción NET de SGML, en la declaración SGML de XML que no hace uso del anexo K de la norma [19], el delimitador NET se declara como '/>', o en el caso de que sí se use el anexo, se utiliza un nuevo separador NESTC (net-enabling start tag close) que se define como '/' y el separador NET, como '>', con lo que se consigue el mismo resultado permitiendo usar NESTC en la etiqueta de inicio solo si va seguida de NET [2].

2.4. MicroXML

En esta sección se estudia el lenguaje MicroXML, una versión reducida de XML que pretende llegar, fundamentalmente, a aquellos usuarios que no han utilizado XML debido a su complejidad. En la introducción al lenguaje se mostrarán los objetivos con los que fue diseñado y se repasarán los conceptos de elemento y atributo. En el resto de apartados, se verá la especificación propuesta para MicroXML. Concretamente, en el apartado dedicado al documento MicroXML se comentará la estructura del documento y, más detalladamente, la sintaxis de elementos y atributos comparándolos con los de XML. El resto de diferencias entre ambos lenguajes se verán en el apartado "Otros aspectos del lenguaje". Finalmente, en el apartado "Analizadores" se probará uno de los analizadores disponibles actualmente escrito por James Clark en JavaScript.

2.4.1. Introducción al lenguaje MicroXML

En los últimos años se están sentando las bases de un nuevo lenguaje basado en XML: MicroXML. En diciembre de 2010, James Clark justifica en su blog [5] la necesidad de una versión reducida de XML. El debate que se abre en la comunidad XML lleva al desarrollo de los primeros borradores sobre MicroXML [11], que describen la estructura de los documentos basados en este lenguaje y la forma de procesarlos. En enero de 2011 John Cowan publica MicroLark, el primer parser para MicroXML, llamado así en honor al parser Lark escrito por Tim Bray para XML. En 2012 se crea un grupo de trabajo (sin carácter oficial) liderado por James Clark, John Cowan y Uche Ogbuji, que publica su primera especificación en octubre de ese mismo año [6]. La gramática propuesta en dicho documento será la que se estudie en los siguientes apartados.

Una versión simplificada de XML podría ser útil para aquellos que no utilizan XML por considerarlo demasido complejo para sus propósitos. MicroXML no es la única alternativa que se está barajando en este sentido. Sin embargo, algunos desarrolladores no encuentran XML suficientemente complejo para requerir una versión simplificada, o piensan que su complejidad ya no supone un impedimento, pues se han desarrollado multitud de herramientas para trabajar con él. James Clark, entre otros, opina que su dificultad sigue teniendo un gran coste para desarrolladores y usuarios y que, por ello, deben buscarse alternativas.

La realidad es que desde el nacimiento de XML han aparecido múltiples versiones que intentaban simplificarlo, como SML en 1999 [22] [31]. Los aspectos más criticados de XML son la forma en que se manejan los errores y la complejidad de los espacios de nombres. Ambas características pretenden mejorarse con MicroXML. El grupo de trabajo de MicroXML destaca, además, las ventajas de este nuevo lenguaje para los recién llegados, que tendrían una formación mucho más sencilla, y para los entornos con limitaciones de memoria o procesamiento, al posibilitar el uso de aplicaciones mucho más ligeras. Señalan también que MicroXML tendría menos problemas de seguridad, pues al carecer de DTDs o entidades, no es necesario acceder a recursos externos para procesarlos. En cualquier caso, MicroXML no pretende sustituir a XML, sino ser una opción más.

Los objetivos iniciales, propuestos por James Clark para este lenguaje, fueron tres:

- Debe ser compatible con XML.

- Debe ser más sencillo trabajar con él tanto para usuarios como para aplicaciones.
- Los documentos deben ser también válidos para HTML5 escogiendo los elementos y atributos apropiados.

El grupo de trabajo de MicroXML amplía estos objetivos con los siguientes:

- La sintaxis de MicroXML debe ser un subconjunto de XML 1.0.
- Debe especificarse una sintaxis y un modelo de datos compatible con XML 1.0.
- Debe ser diseñado para complementar y no sustituir a XML, JSON y HTML.
- Debe soportar el uso de editores de texto para crear documentos.
- Debe soportar UNICODE.
- Debe soportar las necesidades de documentos de contenido mixto.
- Su especificación debe ser tan auto-contenida como sea práctico.

MicroXML es un metalenguaje, una versión reducida de XML, que a su vez, es un subconjunto de SGML y con la que también puede definirse un lenguaje de marcado. Los conceptos fundamentales sobre elementos y atributos siguen siendo los mismos vistos en los apartados 2.2 y 2.3. A modo de introducción, se vuelven a repasar aquí estos conceptos, aunque la descripción detallada puede verse en el apartado 2.2 y las diferencias con los lenguajes anteriores se verán en el siguiente apartado.

El punto de partida es una cierta información que se quiere almacenar en forma de documento MicroXML; para ello, habrá que dividir la información en elementos. Uno de esos elementos puede ser, por ejemplo, el título del documento. El elemento título tendría la siguiente estructura:

```
<titulo>MicroXML</titulo>
```

El elemento está formado por una etiqueta de inicio (<titulo>), un contenido (el propio título, MicroXML) y una etiqueta de fin (</titulo>) (a no ser que se trate de un elemento vacío, como se verá en la siguiente sección). El contenido del elemento puede ser mucho más complejo, pudiendo consistir en otros elementos. De hecho, debe haber un único elemento que contenga a todos los demás (el elemento raíz). Todos los elementos deben estar correctamente anidados, es decir, no se puede dar por terminado un elemento hasta que no terminen los elementos que contiene. Las etiquetas del elemento tampoco tienen porqué limitarse al nombre (identificador) del elemento, pudiendo incluir atributos para aportar información adicional del elemento. En el siguiente ejemplo se incluye el atributo estado al elemento capitulo para indicar que el capítulo se encuentra en fase de borrador.

```
<capitulo estado="borrador">
```

Los ejemplos utilizados son los mismos usados en el capítulo dedicado a XML, pues siguen siendo válidos. En la siguiente sección comenzarán a verse las diferencias.

2.4.2. Documento MicroXML

Un documento MicroXML es una secuencia de caracteres o bytes que siguen las reglas de MicroXML y están codificados en UTF-8. Si un documento sigue la sintaxis de MicroXML, como el de la figura 2.16, debe poder ser parte de un documento SGML y, ser un documento bien formado para XML (no un documento válido, pues puede verse que no tiene DTD).

```
<!--Esto es el primer comentario-->
<informe estado=''borrador''>
<titulo>Primer Informe</titulo>
<p>El informe contiene un título y un párrafo.</p>
</informe>
```

Figura 2.16: Ejemplo de documento MicroXML.

El documento está formado por un único elemento raíz (en el ejemplo es el elemento `informe`) que incluye a los demás (titulo y p) y que, al igual que en XML, deben estar correctamente anidados. Puede comenzar y/o acabar con comentarios y/o espacios en blanco (Donde aparece el carácter 's', aunque se hable de espacios en blanco por simplificar, en realidad se refiere a espacios en blanco, tabuladores y caracteres de nueva línea), como en el ejemplo, sin embargo no puede comenzar con una declaración XML (sólo puede usarse la codificación UTF-8) ni una declaración de tipo de documento, con lo que no se contempla la posibilidad de usar una DTD. Aquí se observa la primera diferencia importante con un documento XML. Se recuerda que el documento XML puede empezar con un prólogo (que debe ser lo primero en aparecer en el documento, si es que existe) que puede contener ambas declaraciones. En MicroXML el documento queda reducido a lo que en XML y SGML se llamó instancia de documento. También se incluye en la especificación la posibilidad de que el documento comience con el byteOrderMark [11].

2.4.2.1. Elementos

La sintaxis de los elementos en MicroXML, igual que en XML, puede ser de dos tipos. Existe una sintaxis para elementos con contenido (como los vistos en la introducción), aunque también puede ser usada por elementos que no lo tengan, y otra sintaxis específica para elementos vacíos. Tanto si tienen contenido como si no lo tienen, a cada elemento se le asigna un identificador, que no es más que un nombre que lo describe.

- **Elementos con contenido**

La sintaxis general, que se usará para elementos con contenido, consiste en una etiqueta de inicio, un contenido, y una etiqueta de fin. Puede tomarse como ejemplo el elemento título que aparece en la Figura 2.16 y que se repite a continuación:

```
<titulo>Primer Informe</titulo>
```

- La etiqueta de inicio (`<titulo>`) comienza con el carácter '<' seguido del identificador de elemento y termina con el carácter '>'. Antes del carácter de cierre es posible que aparezcan espacios en blanco y una lista de atributos (la lista de atributos se describe en la Sección 2.4.2.2).
- La etiqueta de fin de elemento (`</titulo>`) es como la etiqueta de inicio, salvo porque debe comenzar con los caracteres '</' en vez de '<' y porque no lleva atributos.

No se aprecian diferencias entre estas etiquetas y las descritas para XML 1.0. Donde sí empiezan a verse diferencias es en las posibilidades que se tienen para el contenido de un elemento. En el ejemplo comentado, el contenido es una cadena de caracteres (`Primer Informe`). Éste sería el tipo de contenido más sencillo y válido tanto en XML como en MicroXML. En MicroXML un elemento puede contener además a otros elementos, comentarios o referencias de carácter. La sintaxis de los comentarios y las referencias de carácter se ven en los apartados 2.4.3.2 y 2.4.3.1.

Si se compara con la recomendación para XML, puede verse que en ella aparecen dos posibilidades más, las instrucciones de proceso y las secciones CDATA, que no pueden aparecer aquí puesto que ambas herramientas se han eliminado del lenguaje. Además, en XML pueden aparecer referencias a entidades. En MicroXML aunque se han eliminado las entidades, se sigue permitiendo el uso de las referencias de carácter.

A modo de comparación se incluyen la descripciones del contenido de un elemento extraídas de la especificación para MicroXML, la recomendación de XML y la norma para SGML. Así puede verse el gran salto existente entre SGML y MicroXML.

- **Contenido de un elemento según la especificación de MicroXML**

  content ::= (element | comment | dataChar | charRef)*

- **Contenido de un elemento según la recomendación XML 1.0**

  content ::= CharData? ((element | Reference | CDSect | PI | Comment) CharData?)*

- **Contenido de un elemento según la norma ISO 8879 (SGML)**

  content = mixed content | element content | replaceable character data | character data

  mixed content = (data character | element | other content)*

  element content = (element | other content | s)*

  other content = comment declaration | short reference use declaration | link type use declaration | processing instruction | shortref | character reference | general entity reference | marked section declaration | Ee

Figura 2.17: Especificación para el contenido de un elemento en los diferentes lenguajes.

- **Elementos vacíos**

Si un elemento no tiene contenido, puede usarse una sintaxis alternativa, que consiste en una única etiqueta con la misma forma vista para la etiqueta de inicio, pero que acaba

con los caracteres '/>'. Es decir, comienza con el carácter '<' seguido del identificador de elemento y termina con el carácter '/>'. Sería equivalente a usar una etiqueta de inicio y una de fin sin contenido, como en el ejemplo:

</br>

En este caso tampoco se aprecian diferencias con los elementos vacíos en XML.

2.4.2.2. Atributos

Los atributos aportan información adicional sobre el elemento. Se incluyen en la etiqueta de inicio del elemento o en la etiqueta de elemento vacío, en primer lugar aparece el nombre del atributo y, a continuación, el signo '=' seguido del valor entre comillas dobles o simples. En el ejemplo de la figura 2.16 aparece el atributo estado dentro de la etiqueta de inicio del elemento `informe`, que tiene la misma forma que tendría en XML.

```
<informe estado=''borrador''>
```

Como ocurre en XML, puede haber más de un atributo en una misma etiqueta y pueden llevar cualquier orden dentro de ella, con la restricción de que no puede aparecer dos veces el mismo atributo en un mismo elemento. En la especificación también se indica que no puede usarse como nombre de atributo la cadena xmlns. Ésta es una de las mayores diferencias entre ambos lenguajes, pues supone la imposibilidad de usar en MicroXML los espacios de nombres.

Es importante señalar que los espacios en blanco que aparezcan dentro del valor de un atributo no se normalizarán, hecho que puede dar lugar a que dos documentos que sean idénticos para XML no lo sean para MicroXML.

2.4.3. Otros aspectos del lenguaje

2.4.3.1. Caracteres y referencias de carácter

El conjunto de caracteres legales para un documento MicroXML es el UNICODE, sin embargo debemos excluir los llamados códigos de punto subrogados y los no-caracteres. También están prohibidos los caracteres de control, a excepción del espacio en blanco, el tabulador y el carácter de nueva línea (|#x20,#x9 ,#xA). El carácter retorno de carro no se menciona porque, tanto si aparece solo, como si aparece seguido del carácter de nueva línea, ambas posibilidades se reemplazan por un sólo carácter de nueva línea. En XML, además del UNICODE, podía usarse el estándar ISO 10646, que no se incluye en MicroXML.

Al igual que en XML, que un carácter sea legal no quiere decir que pueda usarse en cualquier parte del documento. Todos aquellos caracteres que no formen parte del marcado son considerados caracteres de datos (sin olvidar los valores de los atributos) y el conjunto de caracteres que pueden formar parte de los datos está limitado con el fin de no confundirlos con caracteres de marcado. En MicroXML cualquier carácter legal puede formar parte de los datos de usuario salvo los caracteres '<', '&', '>'. Los dos primeros no pueden formar parte de los datos debido a que son los que indican el comienzo de

una etiqueta y el comienzo de una referencia de carácter, y el carácter '>' se prohibe por compatibilidad con XML. Además, en el caso de los valores de atributo, tampoco pueden aparecer las comillas dobles si se han usado para delimitar el valor del atributo, pero sí pueden utilizarse si se han usado comillas simples para delimitarlo (ocurre lo mismo en el caso opuesto, es decir, si se han usado comillas dobles para delimitar el valor, las comillas simples pueden usarse en el contenido). Si es necesario usar alguno de estos caracteres como parte de los datos, la manera de introducirlos es usando referencias de carácter.

Las referencias de caracteres son útiles, como en XML, para añadir caracteres que no pueden ser introducidos directamente o para añadir a los datos alguno de los caracteres prohibidos ('<', '&' , '>',"' y "").

Existen dos tipos de referencias: numéricas y "nombradas". Las referencias numéricas consisten en sustituir el carácter a introducir por su código hexadecimal precedido por `'&#x'` y seguido por `';'`(por ejemplo, >). En XML puede usarse también una referencia decimal, pero esa posibilidad se elimina en la especificacion de MicroXML. Las referencias "nombradas" permiten hacer referencia a los caracteres de la figura 2.18 de la forma que se indica en la misma.

Carácter	**Referencia**
<	&lt;
>	&gt;
&	&amp;
"	&quot;
'	&apos;

Figura 2.18: Referencias de caracteres en MicroXML.

Los caracteres que forman parte de un identificador de elemento o atributo también están limitados. Los identificadores pueden comenzar, básicamente, por letras o guiones bajos. Cuando se trata de caracteres que van a formar parte del identificador pero no del carácter de comienzo del mismo, se pueden añadir también dígitos (del 0 al 9), guiones y/o puntos. El conjunto completo de caracteres que pueden usarse en los identificadores puede verse en el apartado "Names" de la especificación [6]. Los nombres que comienzan con la cadena "xml" o las combinaciones de estas letras en mayúsculas o minúsculas siguen estando reservados.

2.4.3.2. Comentarios

Los comentarios son idénticos a los vistos en XML, sin embargo, en MicroXML no forman parte del modelo de datos. Dichos comentarios deben comenzar con la cadena `'<!--'` y terminar con `'-->'`.

Pueden añadirse comentarios tanto antes del elemento raíz como dentro de su contenido, siempre y cuando no aparezcan dentro de otra parte del marcado. Por ejemplo, no puede haber un comentario dentro de una etiqueta de inicio o de fin. Por mantener la compatibilidad, los comentarios deben seguir las siguientes reglas:

- Deben empezar con los caracteres '<!--' y terminar con '-->'.
- La cadena '--' no debe aparecer en ningún otro lugar del comentario.
- No puede haber comentarios anidados.

2.4.3.3. Espacios de nombres

Los espacios de nombres han sido eliminados de MicroXML.

Según [42] un espacio de nombres es un conjunto de nombres, identificados por una referencia URI, que se utilizan en documentos XML como tipos de elemento y nombres de atributo. Los espacios de nombres permiten usar elementos y atributos que ya hayan sido definidos y que puedan resultar útiles en otros documentos, en lugar de volver a definirlos. Además permiten el uso de varios elementos o atributos que tengan un mismo identificador, pues podrán diferenciarse por el espacio de nombres al que pertenezcan.

La decisión de eliminar los espacios de nombres debe haber sido una de las decisiones que más tiempo ha tardado en tomar el grupo de trabajo, pues en los primeros borradores aparecía una sección detallada sobre los espacios de nombres en MicroXML. Es una de las simplificaciones del lenguaje que más repercusiones puede tener, y puede suponer un problema para muchos usuarios, pues afecta a la compatibilidad de algunos de sus documentos XML. Según Uche Ogbuji [29] "los espacios de nombres son, con mucho, el concepto más difícil de entender para usuarios y desarrolladores y, además, complican enormemente las especificaciones y el software". Añade además, que se está trabajando en herramientas que ayuden a eliminar y posteriormente reconstruir los espacios de nombres para poder realizar transformaciones de documentos sencillas entre ambas tecnologías. Estas herramientas pueden consultarse en el apartado "Research" de la web del grupo [26].

2.4.3.4. Modelo de datos

Uno de los objetivos de diseño de MicroXML es el siguiente: "Debe especificarse una sintaxis y un modelo de datos compatible con XML 1.0". El grupo de trabajo que desarrolla la especificación considera que el objetivo de poder incluir en ella el modelo de datos es uno de sus objetivos más importantes [29]. Se busca que el modelo de datos sea lo más simple posible y evitar la aparición de modelos de datos distintos, como ocurre en XML, que no incluye el modelo de datos en su especificación.

El modelo de datos de MicroXML contiene sólo tres tipos de datos primitivos: caracteres, listas y mapas. La descripción de cada uno de ellos según la especificación es la siguiente:

- Carácter: es un entero dentro del rango que va de 0 a 0x10FFFF y que representa un código de punto UNICODE.
- Lista: es un tipo estructurado consistente en una lista ordenada con cero o más miebros de cualquier tipo.

- Mapa: es otro tipo estructurado que asocia cero o más claves con un valor. Todas las claves son distintas y tanto las claves como los valores pueden ser de cualquier tipo.

La construcción de alto nivel es el ítem de tipo elemento que contiene tres miembros:

- Ítem de tipo nombre: es una cadena no vacía que consiste en una lista de caracteres.
- Mapa de atributos: es un mapa, puede estar vacío y consiste en:
  - Claves: Son ítems de tipo nombre.
  - Valores: Son cadenas de caracteres.
- Lista de contenidos: es una lista con cero o más miembros. Los miembros pueden ser caracteres u otros ítems de elemento.

Otro de los objetivos de MicroXML es que sea de utilidad para complementar y no para sustituir a XML, JSON y HTML. JSON es un formato de intercambio de datos que se ha hecho popular por su simplicidad, por ello, el grupo de MicroXML ha considerado interesante encontrar la forma de convertir documentos MicroXML a JSON y viceversa. De hecho, el analizador para MicroXML desarrollado por James Clark (que se prueba en el apartado 2.4.3.6) ofrece como salida el modelo de datos con la sintaxis de JSON y en la especificación de MicroXML [6] se incluye también una posible forma de representar el modelo de datos con esta sintaxis. Dicha representación sería la siguiente:

- Un ítem de tipo elemento se representaría como un array JSON.
- Un ítem de tipo nombre se representaría como un string JSON.
- Un mapa de atributos se representaría como un objeto JSON.
- Los valores de un mapa de atributos se representarían como strings JSON.
- La lista de contenidos se representaría como un array JSON.
- Una secuencia de caracteres consecutivos que ocurran en una lista de contenidos se combinaría en un solo string JSON.

2.4.3.5. Manejo de errores

El manejo de errores es uno de los aspectos más criticados de XML. Si un analizador XML encuentra cualquier tipo de error debe detener el análisis inmediatamente. En MicroXML se busca suavizar este comportamiento. Eso no quiere decir que si aparece algún error leve se pueda aceptar el documento como correcto, sino que el analizador no estaría obligado a parar. Podría continuar e informar del error al final, o bien, intentar repararlo, no imponiéndose un tipo concreto de actuación frente a los errores.

2.4.3.6. Analizadores

John Cowan publicó MicroLark en enero de 2011, el primer parser para MicroXML. MicroLark está escrito en Java y cumple las especificaciones del primer borrador editado por él mismo [11]. Está disponible en internet [10] y puede probarse, sin embargo, aún no está actualizado (al menos en el momento en que se ha consultado) para contemplar los cambios posteriores a dicho borrador. Estos cambios sí están recogidos en la primera especificación publicada por el grupo [6], que es la que se ha seguido en este documento y, por ello, se probará uno de los analizadores posteriores. Exiten varios analizadores que pueden consultarse en la página del grupo [25]. James Clark, que es otro de los principales responsables de la aparición de MicroXML, es el desarrollador de uno de ellos. En este apartado se probará este último analizador.

El analizador tiene una versión Java y una versión JavaScript muy sencilla de utilizar (La forma de usarla puede verse más detalladamente en el artículo de Uche Ogbuji [30]). Para probarla basta con descargar los archivos (que pueden encontrarse en el apartado implementaciones de la página del grupo [3]) y, una vez descomprimidos, abrir el archivo test.html. Al hacerlo aparece un cuadro de diálogo donde puede escribirse el archivo MicroXML a analizar.

A continuación se muestran los resultados al analizar un ejemplo de documento MicroXML válido y uno no válido con el parser.

Analizando el ejemplo de la figura 2.16 se obtiene el mensaje de que es un documento correcto y el modelo de datos JSON (figura 2.19).

```
["informe",{"estado":"borrador"},["\n\n",["titulo",{},["Primer
Informe"]],"\n\n",["p",{},["El informe contiene un título y un
párrafo."]],"\n\n"]]
```

Figura 2.19: Modelo de datos JSON del ejemplo de documento MicroXML.

Si se modifica el ejemplo para conseguir un archivo erróneo, eliminando, por ejemplo, una de las etiquetas de fin (figura 2.20), puede comprobarse que el analizador informa del motivo del error: “Parse error: name "informe" in end-tag does not match name "titulo" in start-tag”.

```
<!--Esto es el primer comentario-->
<informe estado=’’borrador’’>
<titulo>Primer Informe
<p>El informe contiene un título y un párrafo.</p>
</informe>
```

Figura 2.20: Ejemplo de documento MicroXML erróneo.

2.5. Conclusiones

El objetivo de este capítulo ha sido el estudio de los lenguajes de marcado. En este estudio, además de revisar el concepto de lenguaje de marcado y su evolución, se han descrito

los lenguajes: SGML, el lenguaje de marcado generalizado estándar, XML, una versión reducida de SGML, y MicroXML, una reducción, a su vez, de XML. Se ha comprobado cómo en XML y MicroXML se han dejado fuera muchas opciones que daban flexibilidad al lenguaje SGML, entre ellas, las características de minimización. El siguiente capítulo se centra en el estudio de dichas características opcionales.

Capítulo 3

Estudio y aplicación de las características de minimización de SGML

Este capítulo se centra en las características de minimización del lenguaje SGML. Se estudia la forma en que deben utilizarse, las dificultades que pueden encontrarse al hacerlo y, además, se aplican al ejemplo utilizado para realizar las medidas del siguiente capítulo. En la Sección 3.2 se utilizan en un ejemplo sencillo, con ello puede entenderse la aplicación correcta de las técnicas antes de pasar a aplicarlas a un archivo mayor y más complejo, como el de la Sección 3.3. En esta última sección se aplica, una a una, cada posible técnica de minimización a los primeros cinco registros de una base de datos con el fin de comparar los resultados obtenidos con cada una y, posteriormente, se determina la combinación de minimizaciones más ventajosa para aplicarla a cincuenta registros. Para trabajar con los archivos generados al minimizar se utilizan un analizador y un editor, por ello se comienza describiendo la instalación de los mismos.

3.1. Instalación del analizador y el editor

El analizador escogido para comprobar la validez de los archivos en los que se aplican las técnicas de minimización es el incluido en OpenSP. OpenSP es utilizado por el servicio de validación del W3C [48] y proviene de uno de los conjuntos de herramientas más populares desarrollados para SGML, SP, de hecho, OpenSP es la versión de SP mantenida por el proyecto OpenJade [33]. SP fue escrito por James Clark para superar las limitaciones de su primer analizador SGMLS, que está basado, a su vez, en uno de los primeros analizadores para SGML publicados, el ARC-SGML de Charles Goldfarb[9].

OpenSP contiene un conjunto de herramientas que facilitan el análisis y la gestión de entidades, entre ellas el analizador Onsgmls, que será utilizado en este proyecto. Onsgmls analiza y valida el documento SGML y muestra una representación de sus ESIS (Element Structure Information Set), la información del documento con la estructura que se haya definido, que será el conjunto de información sobre el que trabajan las aplicaciones que manejan SGML.

Para poder analizar correctamente los archivos minimizados con los que se trabaja, es necesario que el analizador soporte el uso de dichas características y que, además, estén habilitadas en la declaración SGML (como se ve en la Sección 2.2.2.1). La declaración SGML que se usará en los ejemplos es la declaración que usa OpenSP por defecto, donde aparecen habilitadas todas las características de minimización excepto DATATAG (Puede verse en el Anexo D). El motivo de que no aparezca habilitada es que no fue desarrollada por James Clark. Hay autores, como Charles F. Goldfarb [15] y el propio James Clark [4] que consideran que la característica DATATAG (se verá en la Sección 3.2.4) no es recomendable, pues puede ser sustituida de forma más sencilla y flexible usando la característica SHORTREF (Sección 3.2.3). Otros analizadores como el utilizado por OmniMark tampoco soportan DATATAG [32]. De cualquier manera, se verá que no es una de las minimizaciones apropiadas para el ejemplo que se utiliza en este proyecto y, por tanto, no es necesaria para su desarrollo.

En cuanto al editor, podría haberse utilizado cualquier editor sencillo, pero se ha escogido TextPad por ser uno de los muchos que facilitan el trabajo con archivos que contienen etiquetas y por la facilidad con que puede integrarse el analizador como herramienta externa. Esto evita tener que analizar los archivos desde la línea de comandos.

Los archivos se han modificado, analizado y medido en un portátil con sistema operativo Windows XP. En dicho portátil se ha instalado el analizador de la siguiente forma:

1. Se descargan los archivos binarios de la página de distribución de OpenJade, en este caso es la versión 1.5.2:

   http://sourceforge.net/projects/openjade/files/opensp/1.5.2/

2. Se descomprimen los archivos y se extraen en la carpeta deseada, en este caso **C:\opensp**

El análisis puede realizarse directamente en la línea de comandos. Si en esa misma carpeta existe un archivo llamado informe_completo.sgml que quiere analizarse, bastaría con lo siguiente:

- **C:\opensp>onsgmls informe_completo.sgml**

Esta opción no es la que se ha usado habitualmente pues es mucho más cómodo usarlo conjuntamente con el editor.

Para instalar el editor e integrar el analizador se han seguido estos pasos:

1. Se descarga el editor de la página del fabricante.

   http://www.textpad.com/download/index.html

2. Se ejecuta el programa de instalación que incluye.

3. Una vez instalado se accede al menú **Configure** y, dentro del mismo en **Preferences** y **Tools**.

4. Se pulsa el botón **Add Program** y se busca el programa **Onsgmls.exe** instalado en el paso anterior. Después se pulsa **abrir** y **aplicar**.

Para utilizarlo se abre el archivo que se desea analizar y se utiliza el menú **Tools**, **External Tools**, **Onsgmls**. En **Tool Output** puede verse la reconstrucción del documento que hace el analizador y un mensaje sobre el resultado del análisis que bien puede ser de éxito (**Tool completed successfully**) o el mensaje de error correspondiente.

3.2. Estudio de las características de minimización

Las características de minimización descritas para SGML y, que se estudiarán en este apartado, son cinco: OMITTAG, SHORTTAG, SHORTREF, DATATAG y RANK. La primera de ellas, OMITTAG, consiste en eliminar una etiqueta por completo, siempre respetando ciertas reglas para que la eliminación de la etiqueta no afecte a la comprensión de la estructura del documento. Con la siguiente característica, SHORTTAG, se puede reducir el tamaño de una etiqueta, pero sin eliminarla por completo. SHORTREF permite usar referencias abreviadas a entidades de modo que una etiqueta pueda ser sustituída por un carácter, es parecida a DATATAG, que también permite que un carácter actúe como etiqueta, pero en DATATAG no será sólo una etiqueta sino que formará parte de los datos, eliminando así el marcado por completo. Por último RANK permite omitir el nivel de anidamiento de una etiqueta en los casos en que el identificador del elemento incluya este nivel.

A continuación se detalla la forma de utilizar estas características de minimizacion.

3.2.1. OMITTAG

La característica OMITTAG permite omitir ciertas etiquetas de inicio o de fin. La omisión de una etiqueta sólo será posible en aquellos casos en que no sea necesaria para interpretar correctamente el documento. Esto es posible gracias a la DTD, que nos informa previamente de la estructura que sigue el documento y hace innecesarias las etiquetas en aquellos sitios donde puedan inferirse.

Para usar esta característica es necesario especificar, en la declaración de cada uno de los elementos, si se permite la omisión de las etiquetas de comienzo, de las etiquetas de fin o de ambas. Por ejemplo, en declaración del siguiente elemento:

```
<!ELEMENT p - o (#PCDATA|lista)*>
```

Con el guión ‘‘-’’ se indica que no se permite la omisión de la etiqueta de inicio y con la ‘‘o’’ que sí podría usarse esta opción en la etiqueta de fin.

Como se ha comentado anteriormente, una etiqueta podrá omitirse sólo en aquellos casos donde sea posible deducirla por la estructura del documento. En el caso de la etiqueta de fin será posible en tres situaciones:

- Si va seguida por una etiqueta de inicio de otro elemento que no esté incluido en el modelo de contenido del primero.

- Si va seguida por la etiqueta de fin de un elemento que contiene al primero.
- Si se trata de la etiqueta de fin del elemento más externo.

En el primer caso, la razón por la que puede deducirse la etiqueta de fin del elemento, es que si aparece una etiqueta de inicio de otro elemento, que según la DTD no puede estar contenido en el primero, la única opción que se tiene respetando la DTD, es que el primer elemento termine antes de empezar el siguiente.

La segunda situación, ocurre cuando aparece una etiqueta de fin, que no es la que cierra al elemento que se espera, sino que cierra al elemento de un nivel superior en la jerarquía. Si se da por terminado el elemento de nivel superior (que contiene al primer elemento) la única opción siguiendo la DTD, es dar antes por terminados todos los elementos que contiene.

La tercera posibilidad consiste en omitir la etiqueta de fin del elemento raíz. Es la última que debe cerrarse y, por tanto, no puede haber dudas en cuanto a su colocación.

Se experimentará con un ejemplo para verlo más claro. Para ello se escribe el informe de la Figura 3.2, que sigue la DTD de la Figura 3.1, y al que se le aplicará cada una de las posibles minimizaciones. En la Figura 3.3 se muestra el mismo informe al que se le ha aplicado OMITTAG.

```
<!ELEMENT informe            - o   (titulo,seccion+,bibliografia)>
<!ELEMENT seccion            - o   (titulo,p+)>
<!ELEMENT bibliografia       - o   (p+)>
<!ELEMENT (p|titulo)         - o   (#PCDATA)>
```

Figura 3.1: Contenido del archivo informe.dtd.

```
<!DOCTYPE informe SYSTEM "informe.dtd">
<informe>
<titulo>Minimización</titulo>
<seccion>
<titulo>OMITTAG</titulo>
<p>OMITTAG permite omitir completamente algunas etiquetas.</p>
<p>Podemos omitir etiquetas de inicio o de fin.</p>
<p>Debemos asegurarnos de que está habilitada en la declaración SGML.</p>
</seccion>
<bibliografia>
<p>ISO 8879. Information processing -- Text and office systems - Standard
Generalized Markup Language(SGML),1986</p>
</bibliografia>
</informe>
```

Figura 3.2: Contenido del archivo informe_completo.sgml (informe con todas las etiquetas).

```
<!DOCTYPE informe SYSTEM "informe.dtd">
<informe>
<titulo>Minimización
<seccion>
<titulo>OMITTAG
<p>OMITTAG permite omitir completamente algunas etiquetas.
<p>Podemos omitir etiquetas de inicio o de fin.
<p>Debemos asegurarnos de que está habilitada en la declaración SGML.
<bibliografia>
<p>ISO 8879. Information processing -- Text and office systems - Standard
Generalized Markup Language(SGML),1986
```

Figura 3.3: Contenido del archivo informe_omittag.sgml (informe con etiquetas de fin omitidas).

Para hacer uso de OMITTAG se han tenido en cuenta los siguientes aspectos:

- El elemento "título" sólo puede contener caracteres, con lo que abrir cualquier otro elemento después de su etiqueta de inicio debería ser interpretado como fin del título e inicio del nuevo elemento. Gracias a ello puede omitirse la etiqueta de fin del título.

- Un párrafo sólo puede contener datos de tipo carácter. Si dentro de un párrafo se intenta abrir otro, al no ser un elemento permitido, el analizador debería considerar cerrado el primero antes de la etiqueta de inicio del segundo. Esto nos permite omitir algunas etiquetas de fin de párrafo.

- Una sección no puede contener elementos de tipo bibliografía, por lo que al encontrar la etiqueta de inicio de bibliografía, el analizador da por terminada la sección.

- Cuando se da por terminado el informe, su fin debería implicar el cierre de todos los elementos que no hubiesen sido cerrados aún. En este caso son las etiquetas de fin del párrafo de la bibliografía, de la propia bibliografía y del informe, las que pueden omitirse.

Es fundamental, a la hora de aplicar esta y cualquier otra minimización, asegurarse de que no va a ser fuente de errores. Una etiqueta mal situada, que en otras circunstancias supondría un error al analizar el documento, puede ocasionar que un documento sea considerado válido pero que no tenga la estructura que se esperaba. Para comprobar que, efectivamente, una aplicación podría entender la estructura del documento a pesar de la reducción, se ha analizado el documento. El analizador Onsgmls proporciona la salida que se muestra en la Figura 3.4. En ella se comprueba que el documento ha sido considerado válido y que el analizador ha completado las etiquetas que se habían omitido sin ningún problema para identificarlas.

```
(INFORME
(TITULO
-Minimización
)TITULO
(SECCION
(TITULO
-OMITTAG
)TITULO
(P
-OMITTAG permite omitir completamente algunas etiquetas.
)P
(P
-Podemos omitir etiquetas de inicio o de fin.
)P
(P
-Debemos asegurarnos de que está habilitada en la declaración SGML.
)P
)SECCION
(BIBLIOGRAFIA
(P
-ISO 8879. Information processing -- Text and office systems -
Standard\nGeneralized Markup Language(SGML),1986
)P
)BIBLIOGRAFIA
)INFORME
C
Tool completed successfully
```

Figura 3.4: Salida del analizador para informe_omittag.sgml.

Se había comentado anteriormente que, omitir etiquetas sin ceñirse a los casos contemplados en la norma, puede causar resultados inesperados. Puede darse el caso de que el analizador valide el documento y, sin embargo, haya un error en la estructura que pase inadvertido debido a OMITTAG. Es interesante tener en cuenta esta dificultad a la hora de minimizar el documento y ver en qué situaciones habrá que tener precaución con el fin de no cometer este error en los ejemplos posteriores. Para ello se modifica el ejemplo buscando una situación que pueda ser problemática.

En la Figura 3.5 se muestra una nueva DTD y en la Figura 3.6 otro informe. La nueva DTD permite añadir, después del título del informe, varias secciones, párrafos o referencias bibliográficas en cualquier orden.

```
<!ELEMENT informe                   - o  (titulo,(seccion|p|referencia)*)>
<!ELEMENT seccion                   - o  (titulo,p*)>
<!ELEMENT (titulo|p|referencia)     - o  (#PCDATA)>
```

Figura 3.5: Contenido del archivo informe2.dtd.

```
<!DOCTYPE informe SYSTEM "informe2.dtd">
<informe>
<titulo>Minimización
<seccion>
<titulo>OMITTAG
<p>La minimización no debe dificultar la detección de errores.
<p>Podemos consultar el uso correcto de OMITTAG en la norma:
<referencia>ISO 8879. Information processing -- Text and office systems -
Standard Generalized Markup Language(SGML),1986
<p>Los detalles sobre OMITTAG están en el anexo C de la citada norma.
```

Figura 3.6: Contenido del archivo informe_omittag2.sgml (Informe erróneo).

Si se observa la salida del analizador que se ha obtenido para este ejemplo (3.7) no indica que haya ningún error en el marcado. Aparentemente el resultado es correcto, pero el hecho de abrir una referencia donde no debía haberse hecho, ha implicado un cierre del elemento sección, que no era lo que se pretendía. El último párrafo se ha quedado fuera de la sección. Ha sido un error en el uso de OMITTAG y no un error del analizador.

```
(INFORME
(TITULO
-Minimización
)TITULO
(SECCION
(TITULO
-OMITTAG
)TITULO
(P
-La minimización no debe dificultar la detección de errores.
)P
(P
-Podemos consultar el uso correcto de OMITTAG en la norma:
)P
)SECCION
(REFERENCIA
-ISO 8879. Information processing -- Text and office systems -
Standard Generalized Markup Language(SGML),1986
)REFERENCIA
(P
-Los detalles sobre OMITTAG están en el anexo C de la citada norma.
)P
)INFORME
C
Tool completed successfully
```

Figura 3.7: Salida del analizador para informe_omittag2.sgml.

Hasta ahora sólo se ha hablado de la posibilidad de omitir las etiquetas de fin de un

elemento, pero puede hacerse un estudio similar con las etiquetas de comienzo. Podrá omitirse la etiqueta de inicio siempre que el elemento sea requerido por el contexto y el resto de elementos que pudiesen aparecer sean opcionales. Si el elemento tiene atributos que debe especificar o tiene contenido nulo nunca se podría omitir la etiqueta de inicio.

Se usará de nuevo la DTD de la Figura 3.1, pero con una modificación: Ahora se permite la omisión de las etiquetas de inicio (eso no quiere decir que puedan ser realmente omitidas, solo pueden ser omitidas en el caso comentado). En la Figura 3.9 se muestra un informe donde se han omitido, y en la Figura 3.10, el resultado.

```
<!ELEMENT informe            o o   (titulo,seccion+,bibliografia)>
<!ELEMENT seccion            o o   (titulo,p+)>
<!ELEMENT bibliografia       o o   (p+)>
<!ELEMENT (p|titulo)         o o   (#PCDATA)>
```

Figura 3.8: Contenido del archivo informe3.dtd.

```
<!DOCTYPE informe SYSTEM "informe3.dtd">
Minimización
<seccion>
OMITTAG
OMITTAG permite omitir completamente algunas etiquetas.
<p>Podemos omitir etiquetas de inicio o de fin.
<p>Debemos asegurarnos de que está habilitada en la declaración SGML.
<bibliografia>
ISO 8879. Information processing -- Text and office systems - Standard
Generalized Markup Language(SGML),1986
```

Figura 3.9: Contenido del archivo informe_omittag3.sgml (informe erróneo con etiquetas de inicio omitidas).

Según la DTD, lo primero que debe aparecer es la etiqueta de inicio de informe y después de la misma, la etiqueta de inicio de título. Como no hay otra opción pueden omitirse sin problema. Por la misma razón, puede omitirse la etiqueta de inicio del título de sección.

Después del título de sección, lo único que puede aparecer es un párrafo, por ello se ha omitido la etiqueta de inicio del párrafo, sin embargo, existe un problema en este razonamiento: el analizador valida el informe, pero no distingue el título de la primera sección del primer párrafo. Con esto se ve que al combinar la omisión de etiquetas de inicio y de fin, es necesario ser aún más cuidadosos. Al quitar la etiqueta de inicio del párrafo, ya no se puede omitir la etiqueta de fin de título. Añadiendo la etiqueta de fin de título el resultado sería correcto. Habría que decidir entre una de las dos opciones.

```
(INFORME
(TITULO
-Minimización
)TITULO
(SECCION
```

```
(TITULO
-OMITTAG\nOMITTAG permite omitir completamente algunas etiquetas.
)TITULO
(P
-Podemos omitir etiquetas de inicio o de fin.
)P
(P
-Debemos asegurarnos de que está habilitada en la declaración SGML.
)P
)SECCION
(BIBLIOGRAFIA
(P
-ISO 8879. Information processing -- Text and office systems -
Standard Generalized Markup Language(SGML),1986
)P
)BIBLIOGRAFIA
)INFORME
C
Tool completed successfully
```

Figura 3.10: Salida del analizador para el informe erróneo con etiquetas de inicio omitidas.

Si no hubiese otros párrafos, el error obtenido no pasaría inadvertido. Para el informe 3.11 el analizador sí señala un error (Figura 3.12).

```
<!DOCTYPE informe SYSTEM "informe3.dtd">
Minimización
<seccion>
OMITTAG
OMITTAG permite omitir completamente algunas etiquetas.
<bibliografia>
ISO 8879. Information processing -- Text and office systems - Standard
Generalized Markup Language(SGML),1986
```

Figura 3.11: Contenido del archivo informe_omittag4.sgml (segundo informe erróneo con etiquetas de inicio omitidas).

```
C:\sgml\opensp\bin\nsgmls.exe:C:\sgml\spdos\ejemplos\
informe_omittag4.sgml:7:13:E: no start tag specified for
implied empty element "P"
(INFORME
(TITULO
-Minimización
)TITULO
```

```
(SECCION
(TITULO
-OMITTAG\nOMITTAG permite omitir completamente algunas etiquetas.
)TITULO
(P
)P
)SECCION
(BIBLIOGRAFIA
(P
-ISO 8879. Information processing -- Text and office systems -
Standard Generalized Markup Language(SGML),1986
)P
)BIBLIOGRAFIA
)INFORME
Tool completed with exit code 1
```

Figura 3.12: Salida del analizador para informe_omittag4.sgml.

Por último, se añade un ejemplo (Figura 3.13) donde sí sería correcto el uso de OMITTAG tanto en etiquetas de inicio como de fin. Se ha podido omitir la etiqueta de inicio en el informe, en el título del informe, en el título de la sección y en el primer párrafo de la bibliografía.

```
<!DOCTYPE informe SYSTEM "informe3.dtd">
Minimización
<seccion>
OMITTAG
<p>OMITTAG permite omitir completamente algunas etiquetas.
<p>Podemos omitir etiquetas de inicio o de fin.
<p>Debemos asegurarnos de que está habilitada en la declaración SGML.
<bibliografia>
ISO 8879. Information processing -- Text and office systems - Standard
Generalized Markup Language(SGML),1986
```

Figura 3.13: Contenido del archivo informe_omittag5.sgml (informe con etiquetas de inicio y de fin omitidas).

```
(INFORME
(TITULO
-Minimización
)TITULO
(SECCION
(TITULO
-OMITTAG
)TITULO
(P
-OMITTAG permite omitir completamente algunas etiquetas.
)P
(P
-Podemos omitir etiquetas de inicio o de fin.
)P
(P
-Debemos asegurarnos de que está habilitada en la declaración SGML.
)P
)SECCION
(BIBLIOGRAFIA
(P
-ISO 8879. Information processing -- Text and office systems -
Standard Generalized Markup Language(SGML),1986
)P
)BIBLIOGRAFIA
)INFORME
C
Tool completed successfully
```

Figura 3.14: Salida del analizador para informe_omittag5.sgml.

Puede apreciarse que las etiquetas se han reducido notablemente.

3.2.2. SHORTTAG

La característica SHORTTAG permite omitir ciertas partes de una etiqueta, es decir, puede reducirse el tamaño de una etiqueta de varias formas, pero no omitirla por completo. Existen varias posibilidades a la hora de aplicarla y, según las circunstancias, se podrá omitir el separador final de la etiqueta, omitir el identificador (pero no los separadores) u omitir los atributos. Es importante estudiar bajo qué condiciones puede aplicarse cada una de las reducciones, para ello se utilizarán ejemplos que usan la misma DTD escrita para la sección anterior.

3.2.2.1. Omisión del carácter separador TAGC

Puede omitirse el separador final de una etiqueta si va inmediatamente seguida por otra. Se refiere aquí al carácter TAGC (''>''), que marca el final de una etiqueta. Si puede

saberse dónde acaba la etiqueta por otros medios (en este caso la etiqueta debe acabar donde empieza la siguiente, ya que no puede haber una etiqueta dentro de otra), parece razonable que pueda omitirse sin alterar la estructura del documento. En la Figura 3.15 se ha aplicado esta minimización al documento completo. Es una minimización sencilla que no añade gran complejidad para el usuario ni para el parser. En la figura correspondiente a la salida del analizador (Figura 3.16) puede verse que considera correcto el documento.

```
<!DOCTYPE informe SYSTEM "informe.dtd">
<informe
<titulo>Minimización</titulo
<seccion
<titulo>SHORTTAG</titulo
<p>SHORTTAG nos permite omitir ciertas partes de una etiqueta.</p
<p>Nunca podemos omitir la etiqueta por completo.</p
<p>Debemos asegurarnos de que está habilitada en la declaración SGML.</p
</seccion
<bibliografia
<p>ISO 8879. Information processing -- Text and office systems - Standard
Generalized Markup Language(SGML),1986</p
</bibliografia
</informe>
```

Figura 3.15: Contenido del archivo informe_shorttag1.sgml (informe con carácter tagc minimizado).

```
(INFORME
(TITULO
-Minimización
)TITULO
(SECCION
(TITULO
-SHORTTAG
)TITULO
(P
-SHORTTAG nos permite omitir ciertas partes de una etiqueta.
)P
(P
-Nunca podemos omitir la etiqueta por completo.
)P
(P
-Debemos asegurarnos de que está habilitada en la declaración SGML.
)P
)SECCION
(BIBLIOGRAFIA
(P
```

```
-ISO 8879. Information processing -- Text and office systems -
Standard Generalized Markup Language(SGML),1986
)P
)BIBLIOGRAFIA
)INFORME
C
Tool completed successfully
```

Figura 3.16: Salida del analizador para informe_shorttag1.sgml.

3.2.2.2. Omisión del identificador de la etiqueta

Puede omitirse el identificador de la etiqueta si se tienen en cuenta las siguientes reglas:

- En el caso de una etiqueta de fin vacía, el analizador debería completarla con la etiqueta de inicio más cercana.
- En el caso de que sea la etiqueta de inicio la que está vacía hay dos opciones:
  - Si OMITTAG no está habilitado, debería completarla con la etiqueta de fin más cercana y los atributos por defecto.
  - Si OMITTAG está habilitado, debería completarla con la etiqueta de inicio más cercana.

En la Figura 3.17 se ha omitido el identificador de las etiquetas de fin. La aplicación de esta opción al ejemplo no tiene dificultad. Todas las etiquetas de fin son minimizadas pues pueden completarse con la etiqueta de inicio del elemento más cercano que no haya sido aun cerrado. La salida del analizador mostrada en la Figura 3.18 demuestra que el analizador soporta correctamente su uso.

```
<!DOCTYPE informe SYSTEM "informe.dtd">
<informe>
<titulo>Minimización</>
<seccion>
<titulo>SHORTTAG</>
<p>SHORTTAG nos permite omitir ciertas partes de una etiqueta.</>
<p>Nunca podemos omitir la etiqueta por completo.</>
<p>Debemos asegurarnos de que está habilitada en la declaración SGML.</>
</>
<bibliografia>
<p>ISO 8879. Information processing -- Text and office systems - Standard
Generalized Markup Language(SGML),1986</>
</>
</>
```

Figura 3.17: Contenido del archivo informe_shorttag2.sgml (informe con omisión del identificador en las etiquetas de fin).

```
(INFORME
(TITULO
-Minimización
)TITULO
(SECCION
(TITULO
-SHORTTAG
)TITULO
(P
-SHORTTAG nos permite omitir ciertas partes de una etiqueta.
)P
(P
-Nunca podemos omitir la etiqueta por completo.
)P
(P
-Debemos asegurarnos de que está habilitada en la declaración SGML.
)P
)SECCION
(BIBLIOGRAFIA
(P
-ISO 8879. Information processing -- Text and office systems -
Standard Generalized Markup Language(SGML),1986
)P
)BIBLIOGRAFIA
)INFORME
C
Tool completed successfully
```

Figura 3.18: Salida del analizador para informe_shorttag2.sgml.

A la hora de comprobar la posibilidad de omitir los identificadores en las etiquetas de inicio de elemento, se decide hacerlo para el caso en que la característica OMITTAG esté habilitada. El objetivo último es minimizar el archivo tanto como sea posible, por lo que, con el fin de aprovechar al máximo ambas características, se prueba a emplear todas las posibilidades de OMITTAG junto con la omisión del identificador en etiquetas de inicio. Dado que las etiquetas con identificador omitido deben completarse con la etiqueta de inicio más cercana, esta posibilidad sólo puede emplearse con elementos del mismo tipo, como el elemento párrafo del ejemplo. En la Figura 3.19 se muestra el archivo minimizado y en la Figura 3.20 la salida del analizador.

```
<!DOCTYPE informe SYSTEM "informe3.dtd">
Minimización
<seccion>
SHORTTAG
<p>SHORTTAG nos permite omitir ciertas partes de una etiqueta.
<>Nunca podemos omitir la etiqueta por completo.
<>Debemos asegurarnos de que está habilitada en la declaración SGML.
<bibliografia>
ISO 8879. Information processing -- Text and office systems - Standard
Generalized Markup Language(SGML),1986
```

Figura 3.19: Contenido del archivo informe_shorttag3.sgml (informe con omisión de las etiquetas de fin y del identificador de las etiquetas de inicio).

```
(INFORME
(TITULO
-Minimización
)TITULO
(SECCION
(TITULO
-SHORTTAG
)TITULO
(P
-SHORTTAG nos permite omitir ciertas partes de una etiqueta.
)P
(P
-Nunca podemos omitir la etiqueta por completo.
)P
(P
-Debemos asegurarnos de que está habilitada en la declaración SGML.
)P
)SECCION
(BIBLIOGRAFIA
(P
-ISO 8879. Information processing -- Text and office systems -
Standard Generalized Markup Language(SGML),1986
)P
)BIBLIOGRAFIA
)INFORME
C
Tool completed successfully
```

Figura 3.20: Salida del analizador para informe_shorttag3.sgml.

3.2.2.3. Net

Si el separador final de una etiqueta de inicio se sustituye por el separador "null end tag", que suele ser "/", la etiqueta de fin completa puede sustituirse por ese carácter.

Se ve en el ejemplo de la Figura 3.21. Puede complicarse bastante la lectura del documento, lo que puede suponer un inconveniente para el usuario, pero el analizador la interpreta correctamente (Figura 3.22).

```
<!DOCTYPE informe SYSTEM "informe3.dtd">
<informe/
<titulo/Minimización/
<seccion/
<titulo/SHORTTAG/
<p/SHORTTAG nos permite omitir ciertas partes de una etiqueta./
<p/Nunca podemos omitir la etiqueta por completo./
<p/Debemos asegurarnos de que está habilitada en la declaración SGML./
/
<bibliografia/
<p/ISO 8879. Information processing -- Text and office systems - Standard
Generalized Markup Language(SGML),1986/
/
/
```

Figura 3.21: Contenido del archivo informe_shorttag4.sgml (informe en el que se usa net).

```
(INFORME
(TITULO
-Minimización
)TITULO
(SECCION
(TITULO
-SHORTTAG
)TITULO
(P
-SHORTTAG nos permite omitir ciertas partes de una etiqueta.
)P
(P
-Nunca podemos omitir la etiqueta por completo.
)P
(P
-Debemos asegurarnos de que está habilitada en la declaración SGML.
)P
)SECCION
(BIBLIOGRAFIA
```

```
(P
-ISO 8879. Information processing -- Text and office systems -
Standard Generalized Markup Language(SGML),1986
)P
)BIBLIOGRAFIA
)INFORME
C
Tool completed successfully
```

Figura 3.22: Salida del analizador para informe_shorttag4.sgml.

3.2.2.4. Omisión de la lista de atributos

Puede omitirse la lista de atributos o parte de ella si se cumple alguna de las siguientes condiciones:

- Las comillas que suelen delimitar el valor de un atributo pueden omitirse si el valor está compuesto sólo por "name characters". "Name characters" son los caracteres que pueden aparecer en un nombre (letras, dígitos y cualquier otro carácter que indique la sintaxis concreta).
- Si el atributo tiene valores por defecto, o ha sido definido como `#IMPLIED` o como `#CONREF` (Se definen en el apartado 2.2.3.2), puede omitirse por completo.
- Puede utilizarse sólo el valor del atributo y omitir el resto si los posibles valores están incluidos en la dtd.

Se añadirá un atributo a la dtd de los informes con el fin de comprobar que el analizador puede interpretarlo correctamente después de minimizarlo (Figura 3.23). El atributo especifica el soporte en que se almacena el informe (impreso, en línea, en dvd...).

```
<!ELEMENT informe            o o   (titulo,seccion+,bibliografia)>
<!ATTLIST informe
          soporte   CDATA #IMPLIED>
<!ELEMENT seccion            o o   (titulo,p+)>
<!ELEMENT bibliografia       o o   (p+)>
<!ELEMENT (p|titulo)         o o   (#PCDATA)>
```

Figura 3.23: Contenido del archivo informe4.dtd.

```
<!DOCTYPE informe SYSTEM "informe4.dtd">
<informe soporte=''impreso''>
<titulo>Minimización</titulo>
<seccion>
<titulo>OMITTAG</titulo>
<p>OMITTAG permite omitir completamente algunas etiquetas.</p>
<p>Podemos omitir etiquetas de inicio o de fin.</p>
<p>Debemos asegurarnos de que está habilitada en la declaración SGML.</p>
</seccion>
<bibliografia>
<p>ISO 8879. Information processing -- Text and office systems - Standard
Generalized Markup Language(SGML),1986</p>
</bibliografia>
</informe>
```

Figura 3.24: Contenido del archivo informe_completo2.sgml (informe con atributo).

Puede comprobarse que las siguientes líneas serían igualmente válidas en el ejemplo anterior:

```
<informe soporte=''impreso''>
```

```
<informe soporte=impreso>
```

```
<informe>
```

En la última línea el atributo se omite por completo, pues ha sido declarado opcional.

Se modifica ahora la dtd para dar al atributo una lista de posibles valores y un valor por defecto, así podrá comprobarse que, en estos casos, es suficiente con especificar el valor del atributo. Es recomendable omitir el nombre del atributo sólo si está implícito en sus valores, para así evitar la posible ambigüedad [14].

```
<!ELEMENT informe o o (titulo,seccion+,bibliografia)>
<!ATTLIST informe
          soporte (impreso|dvd|internet) impreso>
<!ELEMENT seccion o o (titulo,p+)>
<!ELEMENT bibliografia o o (p+)>
<!ELEMENT (p|titulo) o o (#PCDATA)>
```

Figura 3.25: Contenido del archivo informe5.dtd.

```
<!DOCTYPE informe SYSTEM "informe5.dtd">
<informe dvd>
<titulo>Minimización</titulo>
<seccion>
<titulo>OMITTAG</titulo>
<p>OMITTAG permite omitir completamente algunas etiquetas.</p>
<p>Podemos omitir etiquetas de inicio o de fin.</p>
<p>Debemos asegurarnos de que está habilitada en la declaración SGML.</p>
</seccion>
<bibliografia>
<p>ISO 8879. Information processing -- Text and office systems - Standard
Generalized Markup Language(SGML),1986</p>
</bibliografia>
</informe>
```

Figura 3.26: Contenido del archivo informe_shorttag5.sgml (informe con atributo minimizado).

El analizador interpreta el atributo correctamente (Figura 3.27).

```
ASOPORTE TOKEN DVD
(INFORME
(TITULO
-Minimización
)TITULO
(SECCION
(TITULO
-SHORTTAG
)TITULO
(P
-SHORTTAG nos permite omitir ciertas partes de una etiqueta.
)P
(P
-Nunca podemos omitir la etiqueta por completo.
)P
(P
-Debemos asegurarnos de que está habilitada en la declaración SGML.
)P
)SECCION
(BIBLIOGRAFIA
(P
-ISO 8879. Information processing -- Text and office systems -
Standard Generalized Markup Language(SGML),1986
)P
)BIBLIOGRAFIA
```

```
)INFORME
C
Tool completed successfully
```

Figura 3.27: Salida del analizador para informe_shorttag5.sgml.

3.2.3. SHORTREF

La característica SHORTREF permite usar un carácter o una cadena de caracteres como referencia abreviada a una entidad SGML.

Usando esta característica, podría conseguirse, por ejemplo, que un carácter como el asterisco, fuese sustituído por una etiqueta de inicio de un párrafo o una enumeración. De esta forma, en vez de introducir la etiqueta <p>al principio de cada párrafo, bastaría con introducir un carácter asterisco. Además de la reducción de tamaño del documento y de la comodidad a la hora de escribir, también facilitaría la lectura. Pueden crearse así documentos SGML sin que el usuario sea realmente consciente de estar introduciendo etiquetas.

En el anexo C de la norma [19] se propone usar esta característica para emular a los sitemas WYSWYG (what you see is what you get). En estos sistemas puede editarse el contenido y verlo tal y como quedará una vez impreso, con los saltos de línea, márgenes, etc. Para ello, se insertan códigos en el texto que muestran el efecto deseado y que suelen permanecer ocultos para el usuario. Esto lo hace dependiente de la máquina, que debe ser capaz de entender estos códigos. Con SHORTREF se mantiene la independencia del sistema al interpretarlos como etiquetas SGML.

Para poder usar una referencia corta, se define primero una entidad con la etiqueta que se quiere referenciar. Para el ejemplo que se está comentando se definiría la entidad párrafo:

```
<!ENTITY parrafo "<p>">
```

Después se define la referencia corta que quiere asociarse a esa entidad. Así se obtiene una tabla o mapa de referencias cortas. Si no se usase SHORTREF, cada vez que apareciese una referencia a la entidad párrafo (‘‘&parrafo;’’), sería sustituida por la etiqueta de comienzo de un párrafo, pero aquí, se asocia la entidad párrafo con el carácter asterisco (en una tabla que se ha llamado mapasección), y se vuelve aún más sencillo.

```
<!SHORTREF mapaseccion "*" parrafo>
```

Siempre que ese mapa esté activo, cada aparición de la referencia corta será sustituida por la etiqueta de la entidad. En el ejemplo, para hacer que mapasección sea el mapa activo cuando comience una sección, se usa la siguente línea:

```
<!USEMAP mapaseccion seccion>
```

El mapa seguirá activo en los elementos anidados de la sección, siempre y cuando éstos no tengan sus propios mapas.

En las Figuras 3.28 y 3.29 se muestra el ejemplo completo:

```
<!ENTITY parrafo "<p>">
<!SHORTREF mapaseccion "*" parrafo>
<!USEMAP mapaseccion seccion>
<!ELEMENT informe    - o       (titulo,seccion+,bibliografia)>
<!ELEMENT seccion - o (titulo,p+)>
<!ELEMENT bibliografia - o (p+)>
<!ELEMENT (p|titulo) - o (#PCDATA)>
```

Figura 3.28: Contenido del archivo informe6.dtd.

En el informe de la Figura 3.29 se usa la referencia corta definida en la declaración anterior para comprobar su funcionamiento.

```
<!DOCTYPE informe SYSTEM "informe6.dtd">
<informe>
<titulo>Minimización</titulo>
<seccion>
<titulo>SHORTREF</titulo>
*SHORTREF nos permite usar un carácter o una cadena de caracteres como
referencia abreviada a una entidad SGML.
*Nos permite emular a los sistemas WYSIWYG.
*Nos permite simplificar la escritura de elementos con estructuras
repetitivas como listas o tablas.
</seccion>
<bibliografia>
<p>ISO 8879. Information processing -- Text and office systems - Standard
Generalized Markup Language(SGML),1986</p>
</bibliografia>
</informe>
```

Figura 3.29: Contenido del archivo informe_shortref1.sgml (archivo minimizado con SHORTREF).

En la Figura 3.30 puede comprobarse que el analizador soporta correctamente esta opción.

```
(INFORME
(TITULO
-Minimización
)TITULO
(SECCION
(TITULO
-SHORTREF
)TITULO
(P
-SHORTREF nos permite usar un carácter o una cadena de caracteres como
referencia abreviada a una entidad SGML.
)P
```

```
(P
-Nos permite emular a los sistemas WYSIWYG.
)P
(P
-Nos permite simplificar la escritura de elementos con estructuras
repetitivas como listas o tablas.
)P
)SECCION
(BIBLIOGRAFIA
(P
-ISO 8879. Information processing -- Text and office systems -
Standard Generalized Markup Language(SGML),1986
)P
)BIBLIOGRAFIA
)INFORME
C
Tool completed successfully
```

Figura 3.30: Salida del analizador para informe_shortref1.sgml.

Hay que aclarar, para diferenciar SHORTREF de la característica DATATAG, que el asterisco forma parte del marcado del documento y no de los datos. Si hubiese asteriscos entre los datos de la sección serían interpretados como referencias cortas.

También podría activarse mapasección en una sección concreta, y no en todas ellas. Para ello se elimina de la DTD de la Figura 3.28 la tercera línea, que activaba el mapa, y se modifica el informe tal y como se muestra en la Figura 3.31. La salida del analizador no se añade pues sería idéntica a la anterior.

```
<!DOCTYPE informe SYSTEM "informe6.dtd">
<informe>
<titulo>Minimización</titulo>
<seccion>
<titulo>SHORTREF</titulo>
<!USEMAP mapaseccion>
*SHORTREF nos permite usar un carácter o una cadena de caracteres como
referencia abreviada a una entidad SGML.
*Nos permite emular a los sistemas WYSIWYG.
*Nos permite simplificar la escritura de elementos con estructuras
repetitivas como listas o tablas.
</seccion>
<bibliografia>
<p>ISO 8879. Information processing -- Text and office systems - Standard
Generalized Markup Language(SGML),1986</p>
</bibliografia>
</informe>
```

Figura 3.31: Contenido del archivo informe_shortref2.sgml (segundo informe minimizado con SHORTREF).

Para desactivar todos los mapas se usaría la línea: `<!USEMAP #EMPTY>`.

Muchos de los signos de puntuación pueden usarse como referencias cortas. También pueden usarse como tales algunas secuencias de control, por ejemplo, en vez de usar un asterisco como delimitador de inicio de párrafo, podría usarse de la misma manera una línea en blanco (`&#RS;&#RE;`). Cada aparición de una línea en blanco sería equivalente a la etiqueta `<p>`.

En la sintaxis concreta se especifican los caracteres y secuencias de control que pueden ser usados como referencias cortas. Se incluyen como anexo (Anexo C).

3.2.4. DATATAG

La característica DATATAG permite definir caracteres que serán interpretados a la vez como etiquetas y como datos.

Su uso se basa en la definición de un patrón que seguirán los datos y que permitirá al analizador identificar el marcado del documento. El analizador OpenSP no soporta esta característica. Para comprobarlo, se usará parte de un ejemplo del anexo C de la norma ISO 8879, que debería servir para separar los párrafos en oraciones, es decir, lo que se busca es poder escribir un párrafo completo y que el analizador sea capaz de introducir etiquetas para cada una de sus oraciones.

Lo primero es definir cómo es el "data tag group". Si se tiene un párrafo como el siguiente:

```
<p>Primera oración. Segunda oración. Tercera oración.</p>
```

El analizador debería ser capaz de interpretar que el párrafo está compuesto por tres oraciones y tres "paradas" que delimitan esas oraciones. El patrón para el elemento párrafo sería:

```
([oracion,%stop;]+)
```

Aquí "stop" es una referencia a una entidad que se usa para definir los elementos que pueden delimitar una oración, como por ejemplo un punto y un espacio. El ejemplo completo se ve en las Figuras 3.32 y 3.33. La Figura 3.37 es la salida del analizador, que demuestra que, aunque da por correcta la estructura del documento, no es capaz de distinguir las dos oraciones dentro del párrafo.

```
<!ENTITY% stop '( ".&#RE;" | ". " | ".)&#RE;" | ".)" |
                  "?&#RE;" | "? " | "?)&#RE;" | "?)" |
                  "!&#RE;" | "! " | "!)&#RE;" | "!)" )'>
<!ELEMENT informe - o (titulo,seccion+,bibliografia)>
<!ELEMENT seccion - o (titulo,p+)>
<!ELEMENT bibliografia - o (p+)>
<!ELEMENT titulo - o (#PCDATA)>
<!ELEMENT p - o ([oracion,%stop;]+)>
<!ELEMENT oracion o o (#PCDATA)>
```

Figura 3.32: Contenido del archivo informe7.dtd.

```
<!DOCTYPE informe SYSTEM "informe7.dtd">
<informe>
<titulo>Minimización</titulo>
<seccion>
<titulo>DATATAG</titulo>
<p>
DATATAG nos permite definir caracteres que serán interpretados a la vez
como etiquetas y como datos.
Se quiere comprobar que el analizador OpenSP no soporta esta
característica.
</p>
</seccion>
<bibliografia>
<p>ISO 8879. Information processing -- Text and office systems - Standard
Generalized Markup Language(SGML),1986</p>
</bibliografia>
</informe>
```

Figura 3.33: Contenido del archivo informe_datatag.sgml (informe minimizado usando Datatag).

```
(INFORME
(TITULO
-Minimización
)TITULO
(SECCION
(TITULO
-SHORTREF
)TITULO
(P
(ORACION
-DATATAG nos permite definir caracteres que serán interpretados a la vez
como etiquetas y como datos. Se quiere comprobar que el analizador SP no
soporta esta característica.
)ORACION
)P
)SECCION
(BIBLIOGRAFIA
(P
(ORACION
-ISO 8879. Information processing -- Text and office systems -
Standard Generalized Markup Language(SGML),1986
)ORACION
)P
```

```
)BIBLIOGRAFIA

)INFORME

C

Tool completed successfully
```

Figura 3.34: Salida del analizador para informe_datatag1.sgml.

Las ventajas de esta característica serían muchas a la hora de almacenar documentos que siguen un patrón, por ejemplo, en una base de datos. Se puede eliminar por completo la necesidad de introducir etiquetas en esos casos, con lo que, además de ahorro de memoria, sería lo más simple para el usuario. La desventaja sería la complejidad de las aplicaciones que manejen los datos.

3.2.5. RANK

La característica RANK permite omitir el nivel de anidamiento de algunas etiquetas.

Si se quiere especificar el nivel de anidamiento de varios elementos, es necesario dividir su identificador en dos partes: una raíz y un sufijo. El sufijo debe ser un número que indique el nivel. Con RANK sólo habrá que especificar el sufijo para el primero de los elementos del mismo nivel, y el analizador considerará todos los siguientes pertenecientes al mismo mientras no se indique otro explícitamente.

El siguiente ejemplo muestra el uso de la característica en las Figuras 3.35 y 3.36, y la confirmación de que es soportada por el analizador en la Figura 3.37.

```
<!ELEMENT informe - o (titulo,seccion+,bibliografia)>
<!ELEMENT seccion - o (titulo,p1+)>
<!ELEMENT bibliografia - o (p1+)>
<!ELEMENT titulo - o (#PCDATA)>
<!ELEMENT p 1 - o (#PCDATA,p2*)>
<!ELEMENT p 2 - o (#PCDATA)>
```

Figura 3.35: Contenido del archivo informe8.dtd.

```
<!DOCTYPE informe SYSTEM "informe8.dtd">
<informe>
<titulo>Minimización</titulo>
<seccion>
<titulo>RANK</titulo>
<p1>RANK nos permite determinar el nivel de anidamiento sin especificarlo
explícitamente.
<p>Para poder usarlo debemos seguir los siguientes pasos a la hora de
declararlo:
<p2>Debemos dividir el identificador del elemento en raíz y sufijo.
<p>El sufijo debe ser un número.
<p1>Sólo necesitamos especificar el nivel en el primero de los elementos
que escribamos mientras queramos conservar el mismo.
</seccion>
<bibliografia>
<p1>ISO 8879. Information processing -- Text and office systems -
Standard Generalized Markup Language(SGML),1986
</bibliografia>
</informe>
```

Figura 3.36: Contenido del archivo informe_rank1.sgml (informe minimizado usando Rank).

```
(INFORME
(TITULO
-Minimización
)TITULO
(SECCION
(TITULO
-RANK
)TITULO
(P1
-RANK nos permite determinar el nivel de anidamiento sin especificarlo
explícitamente.
)P1
(P1
-Para poder usarlo debemos seguir los siguientes pasos a la hora de
declararlo:
(P2
-Debemos dividir el identificador del elemento en raíz y sufijo.
)P2
(P2
-El sufijo debe ser un número.
)P2
)P1
```

```
(P1
-Sólo necesitamos especificar el nivel en el primero de los elementos que
escribamos mientras queramos conservar el mismo.
)P1
)SECCION
(BIBLIOGRAFIA
(P1
-ISO 8879. Information processing -- Text and office systems -
Standard Generalized Markup Language(SGML),1986
)P1
)BIBLIOGRAFIA
)INFORME
C
Tool completed successfully
```

Figura 3.37: Salida del analizador para informe_rank1.sgml.

3.3. Aplicación de las características de minimización

Una vez estudiadas las características de minimización se aplican al documento utilizado para las medidas. Se ha escogido uno de los tipos de documentos que más pueden beneficiarse de este tipo de reducción, los que almacenan información de bases de datos. En este tipo de documento las etiquetas representan un porcentaje muy elevado del tamaño del archivo.

El ejemplo consiste en un pequeño fragmento extraído de una gran base de datos creada con propósitos educativos por Gio Wiederhold [39]. La base de datos viene dada como siete archivos XML con sus correspondientes DTDs. Se usará sólo un fragmento de uno de los archivos que contiene la información correspondiente a cincuenta registros y su correspondiente DTD. Dicha DTD se muestra en la Sección 3.3.1, donde también se aclara la importancia de la misma a la hora de poder aplicar o no, algunos tipos de minimización. Antes de minimizar el documento completo extraído de la base de datos, se decide hacer una comparación de las distintas técnicas de minimización. En esta comparación se usarán sólo los cinco primeros registros. El objetivo de la misma, además de estudiar por separado las posibles ventajas de cada técnica, es determinar la forma de aplicar una combinación de minimizaciones que dé como resultado el archivo más reducido posible. Dicha combinación de técnicas es la que posteriormente se aplica al documento con cincuenta registros.

Todos los archivos minimizados han sido analizados con OpenSP y fueron considerados válidos. No se incluyen las salidas del analizador porque son extensas y alargarían el texto sin aportar más información.

3.3.1. DTD utilizada

La DTD que sigue un archivo SGML debe estudiarse detenidamente para determinar qué características de minimización van a poder utilizarse y en qué elementos. Se debe prestar atención por un lado, a los caracteres de la declaración de elemento que indican si las etiquetas de inicio y/o de fin de elemento pueden omitirse, y por otro, al tipo de contenido de los elementos, el orden en que pueden aparecer y los indicadores de ocurrencia.

Los caracteres usados para indicar la posibilidad de omisión de etiquetas son dos. El carácter '`o`' indica que una etiqueta podría omitirse, y el carácter '`-`' que no debe omitirse, donde el primero en aparecer corresponde a la etiqueta de comienzo, y el segundo a la etiqueta de fin. El que aparezca el carácter '`o`' en la declaración de un elemento quiere decir que se permite su omisión, pero sólo puede omitirse realmente si el elemento se encuentra, además, en uno de los casos contemplados por la norma de SGML. De ahí viene la importancia de estudiar la estructura del documento antes de minimizar. Por ejemplo, si en la DTD se observa que cada registro tiene un elemento concreto que debe aparecer necesariamente y debe hacerlo en primer lugar, es un elemento requerido por el contexto, y su etiqueta de inicio es prescindible (como se ve en la Sección 3.2.1), pues puede ser determinada gracias a la DTD. Sin embargo, si dicho elemento pudiese aparecer más de una vez , no podría omitirse la etiqueta de inicio en el resto de apariciones porque no se tendría la certeza de que fuese a aparecer.

Al aplicar las normas estudiadas en la Sección 3.2 a una DTD concreta, pueden deducirse una serie de pautas a seguir que permiten aplicar las minimizaciones de una forma prácticamente automática. Para determinar esas pautas se empieza estudiando la estructura que deben tener los documentos que sigan dicha DTD.

Se debe comentar, que aunque el ejemplo se haya extraído de la base de datos citada anteriormente, los archivos con los que se trabaja, que son los que aquí se muestran, no son totalmente idénticos a los encontrados. Se han hecho pequeñas modificaciones. Algunas modificaciones eran necesarias, como añadir los caracteres que en SGML indican la posibilidad, o no, de omitir etiquetas (la DTD no los tenía pues era para documentos XML) y cambiar algunos indicadores de ocurrencia para subsanar algunos errores encontrados, y otras se han hecho para mejorar lo que se quiere ilustrar, como eliminar algunos elementos no usados en los registros que se consideran (ya que se trabaja sólo con un extracto de la base de datos) y cambiar algunos elementos por atributos. Por otro lado, se quiere señalar que hay elementos en cuya declaración no se ha permitido la omisión de alguna de sus etiquetas. Esto es así porque se decide partir de una DTD en la que no se permite ninguna omisión y, a medida que se estudia, se va permitiendo la omisión de aquellas etiquetas que se sabe que, efectivamente, serán omitidas. Así se evitan y detectan posibles errores más fácilmente.

La DTD resultante se muestra en la Figura 3.38. Puede verse que permite almacenar información sobre películas. Cada documento que la siga, debe tener un elemento raíz llamado `movies`, el cual debe contener subelementos de tipo `film`, cada uno de ellos representa un registro. Los registros deben contener necesariamente información sobre el título (`t`), el año en que se termina la película (`year`) y la lista de directores (`dirs`). Opcionalmente puede incluir la fecha del estreno (`date`), datos sobre los productores (`prods`), los estudios donde se ha rodado (`studios`), los procesos usados (`prcs`), categorías (`cats`), premios (`awards`), localizaciones (`loc`), periodos de tiempo (`period`),

gente que colabora en la película (`people`) y posibles errores que deben ser verificados (`error`). También son opcionales, aunque pueden aparecer cualquier número de veces los títulos alternativos (`alts`) y notas sobre la película (`notes`).

Los elementos mencionados deben aparecer necesariamente en el orden indicado en el modelo de contenido del elemento `film`. Esto viene determinado por el conector utilizado entre los elementos, que en este caso es el carácter ','. Dichos elementos pueden contener datos, que en este caso están representados por la palabra reservada `#PCDATA` (caracteres de datos que pueden ser analizados en busca de marcado), o bien pueden contener otros elementos con modelos de contenidos muy similares al del elemento `film`, pues son grupos de elementos que sólo usan el conector ',' y muchos son opcionales.

Hay pocos atributos. Sólo tienen atributos los elementos `film`, `dir` y `prod`. Los valores de los atributos de los elementos `dir` y `prod` están indicados en la DTD, eso significa que a la hora de especificarlos bastará con poner el valor del atributo después del nombre del elemento y podrá omitirse el resto de la especificación, incluso podrá omitirse por completo si el valor coincide con el valor por defecto especificado en la DTD.

A continuación se muestra la DTD utilizada y en las siguientes secciones se determinan las pautas concretas que se siguen al aplicar cada tipo de minimización.

```
<!ELEMENT movies o o (film*)>

<!ELEMENT film - o (t, alts*, year, date?, dirs, prods?, studios?, prcs?,
cats?, awards?, loc?, period?, people?, notes*, error?)>

<!ATTLIST film fid ID #REQUIRED>

<!ELEMENT t o o (#PCDATA)>

<!– The film's title –>

<!ELEMENT alts - o (alt+) >

<!– Alternate titles. –>

<!ELEMENT alt o o (altt, altwhy?) >

<!– Alternate titles. also used for multi-part movies –>

<!ELEMENT altt o o (#PCDATA)>

<!– Alternate titles. –>

<!ELEMENT altwhy - o (#PCDATA)>

<!– Reason for alternate titles. date, or country name, or part–>

<!ELEMENT year o o (#PCDATA)>

<!- Year the movie was completed. –>

<!ELEMENT date - - (#PCDATA)>

<!– Date first shown, mainly used for Hitchcock TV shows –>

<!ELEMENT dirs o o (dir+, diraward*, dirnote?)>

<!– Film-specific director information. –>

<!ELEMENT dir - o (dirn) >
```

<– director entry –>

<!ATTLIST dir dirk (R|A|N) N>

<!ELEMENT dirn o o (#PCDATA)>

<!– The first name should be redundant with dirn–>

<!ELEMENT diraward - - (#PCDATA)>

<!– awards given the director for this film . –>

<!ELEMENT dirnote - - (#PCDATA)>

<!– note about the direction, rare. –>

<!ELEMENT prods - o (prod*, prodnote?)>

<!– Producer(s) of the movie –>

<!ELEMENT prod - o (pname)>

<!– Multiple producers are permitted and common.–>

<!– currently prdname indicates not the canonical name –>

<!ATTLIST prod prodk (R|A|C|O|N|X|M|S) S>

<!– If value is R the producer's name in pname is canonical and in People.xml file –>

<!– if value is A, the producer is an actor in actors.xml–>

<!– if value is C, the producer is a cinematographer in people –>

<!– if value is O indicates Official Agency; –>

<!– if value is N, it is not canonical now, and –>

<!– X means it has not been processed as R, A, or N –>

<!– M indicates more, unknown producers –>

<!– if value is S, then the spelling is uncertain. No reference to people can be expected. –>

<!ELEMENT pname o o (#PCDATA)>

<!– Producers name, if prodk is not R, then it' s the full name –>

<!ELEMENT prodnote - o (#PCDATA)>

<!– Note about the producerers. Rare –>

<!ELEMENT studios - o (studio+, studioloc*, distributor*)>

<– Studio(s) where the movie was filmed–>

<!ELEMENT studio o o (#PCDATA)>

<!ELEMENT studioloc - o (#PCDATA)>

<!ELEMENT distributor - o (#PCDATA)>

<!ELEMENT prcs - o (prc+, prctext*, length?, lang?) >

<!– Process used to make the movie, no contents if unknown –>

<!ELEMENT prc o o (#PCDATA)>

<!–code for process(e.g. black and white as 'bnw', col). Color processes can be specific–>

<!ELEMENT prctext - o (#PCDATA)>

<!– rare, additional comments if process is unusual. Also used for delayed, re-releases–>

<!ELEMENT length - - (#PCDATA)>

<!– unusual length, as less than 1 hour, more than 3 hours. –>

<!ELEMENT lang - - (#PCDATA)>

<!– original language(s), if not English, uses Country codes.. –>

<!ELEMENT cats - o (cat+, cattext?) >

<!– Categories assigned to this film –>

<!ELEMENT cat o o (#PCDATA)>

<!– Category of the film, coded. –>

<!ELEMENT cattext - - (#PCDATA)>

<!– Category of the film, textual –>

<!ELEMENT awards - o (aw+) >

<!– Award information –>

<!ELEMENT aw o o (awtype, awattr?, awref?)>

<!– specific award –>

<!ELEMENT awtype o o (#PCDATA)>

<!– specific award type. Coded with award entry. –>

<!ELEMENT awattr - o (#PCDATA)>

<!– Comment: Notes about award level, VIP, or recognition of film. –>

<!ELEMENT awref - o (#PCDATA)>

<!–reference for award citation, common with VIP.. –>

<!ELEMENT loc - o (site+)>

<!–Locations where the film plays. –>

<!ELEMENT site o o (filmedat?, sitename?, sitedes?, siteclass?, siteat?, siteplace?) >

<!ELEMENT filmedat - - (#PCDATA)>

<!– flag indicates site is actual location of filming, if different and significant –>

<!ELEMENT sitename - o (#PCDATA)>

<!– name of location, may be fictional –>

<!ELEMENT sitetype - - (#PCDATA)>

<!– type of site, not coded. –>

<!ELEMENT sitedes - o (#PCDATA)>

<!– description of site, may use codes given in codes.rdf. –>

<!ELEMENT siteclass - o (#PCDATA)>

<!– type of location, uses codes given in codes.rdf. –>

<!ELEMENT siteat - o (#PCDATA)>

<!– additional attributes of the place of the site, as East, West, or areas, or cities –>

<!ELEMENT siteplace - o (#PCDATA)>

<!– used for actual relevant geographic areas –>

<!ELEMENT period - o (#PCDATA)>

<!– If the period of the film is significant it is given as [[dd]mmm]yypp[AD,BC])–>

<!ELEMENT people - o (authors*, writers?, visuals?, choreographers?, cingraphs?, composers?, editors?)>

<!– Information about people involved in this film.–>

<!ELEMENT authors - o (names+, bt?, pawards*) >

<!– names of authors –>

<!ELEMENT writers - o (names+, pawards*) >

<!– names of writers –>

<!ELEMENT visuals - o (names+, fnote?, pawards*) >

<!– creators of visuals –>

<!ELEMENT choreographers - - (names+, bt?, fnote?, pawards*) >

<!– choreographers –>

<!ELEMENT cingraphs - o (names+, fnote?, pawards*)>

<!– cinematographers –>

<!ELEMENT composers - o (names+, bt?, pawards*) >

<!– music composer or performer –>

<!ELEMENT editors - - (names+, pawards*) >

<!ELEMENT names o o (kname?, name*, morenames?)>

<!– names of authors –>

<!ELEMENT kname - o (#PCDATA)>

<!– Name of person, appears in people file. . –>

<!ELEMENT name - o (#PCDATA)>

<!– Name of person, may appear in people file. –>

<!– fix by having new entry type –>

<!ELEMENT morenames - - (#PCDATA)>

<!– Placeholder for more names, missing. –>

<!ELEMENT bt - o (#PCDATA)>

<!– booktitle or title of ballet. Only one book or ballet per film. –>

<!ELEMENT pawards - o (paw+, pawattr?) >

<!– Awards given to all authors for work in this film. –>

<!ELEMENT paw o o (#PCDATA)>

<!– Award type, uses award codes. –>

!ELEMENT pawattr - - (#PCDATA)>

<!– information about this award. –>

<!ELEMENT fnote - - (#PCDATA)>

<!ELEMENT notes - o (crossref?, rating?, money?, facts*, source?) >

<!– A variety of notes is kept. –>

<!ELEMENT crossref - - (reftype?, refdest, refto) >

<!– crossrefers from film –>

<!ELEMENT reftype - - (#PCDATA) >

<!– Motivation –>

<!ELEMENT refdest - - (#PCDATA)>

<!– destination code, URI or local Film, Director, Studio, Actor –>

<!ELEMENT refto - - (#PCDATA)>

<!– URI or referenced object name –>

<!ELEMENT rating - - (#PCDATA)>

<!– rating codes –>

<!ELEMENT money - o (budget?, cost?, inc?, profit?, moneynotes*) >

<!– Information about movie finances –>

<!ELEMENT budget - - (#PCDATA)>

<!–Planned cost of film, by default in US dollars –>

<!ELEMENT cost - o (#PCDATA)>

<!–Approximate cost of film, by default in US dollars –>

<!ELEMENT inc - - (#PCDATA)>

<!–Approximate income for film, default in US dollars –>

<!ELEMENT profit - o (#PCDATA)>

<!–Approximate profit, i.e., inc - cost, US dollars –>

<!ELEMENT moneynotes - o (#PCDATA)>

<!– Extraordinary financial notes, including currency change –>

<!ELEMENT facts - o (#PCDATA)>

<!– Other information about the movie. fields are comma delimited for now –>

<!ELEMENT source - o (by?, seen?, vt?) >

<!– Information about the sources for information –>

<!ELEMENT by - - (#PCDATA)>

<!– Person or unusual reference, if missing likely to be gio or student –>

<!ELEMENT seen - o (#PCDATA)>

<!– date(s) that movie was seen –>

<!ELEMENT vt - o (#PCDATA)>

<!– videotape designation for that movie –>

<!ELEMENT error - o (#PCDATA)>

<!– Documents a variety of possible errors to be verified –>

Figura 3.38: DTD utilizada en el documento que se estudia (dtd_modificada.dtd).

3.3.2. Documento sin minimizar

En esta sección se muestra un fragmento del archivo SGML que se quiere minimizar. Dicho fragmento contiene los primeros cinco registros de la base de datos. El documento completo, con cincuenta registros, se incluye en el Anexo A.1 debido a su longitud. Es interesante fijarse en la cantidad de etiquetas que han sido necesarias en cada registro para almacenar unos pocos datos sobre una película, las etiquetas representan más de un 80 % del tamaño del documento (las medidas se ven en la sección 4.2) y eso significa que si las etiquetas pueden reducirse notablemente, también será importante la reducción del tamaño de archivo.

Antes de minimizar el documento completo, se pretende hacer una comparación de las distintas técnicas de minimización y determinar la combinación de las mismas que daría como resultado el archivo más reducido posible. Para hacer esta comparación se utilizarán archivos más manejables, un archivo con un registro y otro con cinco registros (que es el incluido en la Figura 3.39), a los que se aplicarán las técnicas de minimización por separado. Estos dos archivos y sus correspondientes archivos minimizados se utilizarán en el Capítulo 4 para comprobar la efectividad de cada tipo de minimización, reduciendo el tamaño de los archivos y las variaciones con el número de registros considerados. El documento sin minimizar con un solo registro no se incluye pues, simplemente, es el resultado de extraer el primer registro del documento que se muestra a continuación.

```
<!DOCTYPE movies SYSTEM "dtd_modificada.dtd">
<movies>
 <film fid="H1">
```

```
 <t>Always Tell Your Wife</t>
 <year>1922</year>
 <dirs>
  <dir dirk="R">
   <dirn>Se.Hicks</dirn>
  </dir>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Lasky</pname>
  </prod>
 </prods>
 <studios>
  <studio>Famous</studio>
 </studios>
 <prcs>
  <prc>sbw</prc>
 </prcs>
 <cats>
  <cat>Dram</cat>
 </cats>
</film>

<film fid="H2">
 <t>Number Thirteen</t>
 <year>1922</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Hitchcock</pname>
  </prod>
 </prods>
 <studios>
  <studio>Islington</studio>
  <distributor>Famous</distributor>
 </studios>
 <prcs>
  <prc>sbw</prc>
  <prctext>unfinished</prctext>
 </prcs>
```

```
</film>

<film fid="H3">
 <t>Woman to Woman</t>
 <year>1922</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
  <dir dirk="N">
   <dirn>Cutts</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Balcon</pname>
  </prod>
 </prods>
 <studios>
  <studio>B-S-F</studio>
  <distributor>Wardour</distributor>
 </studios>
 <prcs>
  <prc>sbw</prc>
 </prcs>
 <cats>
  <cat>Dram</cat>
 </cats>
 <loc>
  <site>
   <siteplace>GB</siteplace>
  </site>
 </loc>
 <error>same(GCt27), y(1926)</error>
</film>

<film fid="H4">
 <t>The Passionate Adventure</t>
 <year>1924</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
  <dir dirk="N">
   <dirn>Cutts</dirn>
  </dir>
 </dirs>
```

```
 <prods>
  <prod prodk="R">
   <pname>Balcon</pname>
  </prod>
 </prods>
 <studios>
  <studio>Gainsborough</studio>
  <distributor>GaumontD</distributor>
 </studios>
 <prcs>
  <prc>sbw</prc>
 </prcs>
</film>

<film fid="H5">
 <t>The Blackguard</t>
 <year>1925</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
  <dir dirk="N">
   <dirn>Cutts</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Balcon</pname>
  </prod>
 </prods>
 <studios>
  <studio>UFA</studio>
  <distributor>Wardour</distributor>
 </studios>
 <prcs>
  <prc>sbw</prc>
 </prcs>
</film>
</movies>
```

Figura 3.39: Documento sin minimizar con cinco registros (registro5.sgml).

3.3.3. Aplicación de OMITTAG

Tal y como se vio en la sección 3.2.1, existen dos posibilidades al aplicar OMITTAG, pueden omitirse las etiquetas de inicio, o bien las etiquetas de fin. Además, ambas

posibilidades pueden usarse conjuntamente, aunque ya se vio en esa misma sección que había que hacerlo con especial cuidado, pues podía ser fuente de errores. No basta con comprobar que el analizador valida el documento, sino que debe ser capaz de interpretar correctamente su estructura.

3.3.3.1. Omisión de las etiquetas de fin

Hay tres situaciones en las que está permitido omitir una etiqueta de fin:

1. Si va seguida por una etiqueta de inicio de otro elemento que no esté incluido en el modelo de contenido del primero.
2. Si va seguida por la etiqueta de fin de un elemento que contiene al primero.
3. Si se trata de la etiqueta de fin del elemento más externo.

En el caso del documento considerado, muchos de los elementos contienen únicamente datos de usuario, para este tipo de elementos puede omitirse la etiqueta de fin, pues cuando el analizador encuentre la etiqueta de inicio del siguiente elemento (que claramente no pertenece al modelo de contenido del primero puesto que sólo puede contener datos) dará por terminado el primero. Sería el primer supuesto en que la norma permite la omisión de una etiqueta. Por ejemplo, en el caso del elemento `t`, que contiene la información del título, cuando el analizador encuentra la etiqueta de inicio del elemento `year,` da por terminado el elemento.

Si el elemento cuya etiqueta de fin se desea omitir tiene subelementos, es necesario asegurarse de que, esos mismos elementos que forman parte de su modelo de contenido, no puedan aparecer como elementos externos al primero, en este caso se continuaría en el primero de los supuestos. En el ejemplo considerado no puede darse esta situación si no se ha omitido ninguna etiqueta de inicio. Si se ha hecho sí habría que prestar atención, pues hay subelementos con el mismo nombre que pueden estar contenidos en elementos distintos. De aquí puede concluirse que, en este caso concreto, las etiquetas de fin que vayan seguidas por etiquetas de inicio de otros elementos pueden ser omitidas. Esto permitiría omitir todas las etiquetas de fin excepto las últimas cuatro etiquetas del documento.

La última etiqueta del documento, la del elemento `movies`, puede ser omitida por ser la etiqueta de fin del elemento más externo. En cuanto a las etiquetas restantes, puede considerarse que cumplen el segundo de los supuestos. Al ir seguidas por la etiqueta de fin del elemento `movies`, que contiene a todos los demás, cuando el analizador encuentra o infiere el final de movies, da por finalizados todos los elementos contenidos en él que no hayan sido cerrados.

Puede parecer complicado, sin embargo, este análisis sólo se ha hecho una vez, pues sólo ha sido necesario fijarse en la DTD, y una vez obtenida la conclusión de que todas las etiquetas de fin pueden omitirse, no tiene dificultad aplicarlo a los registros, podría realizarlo una aplicación sencilla, y el resultado podría ser interesante pues se han eliminado el 50 % de las etiquetas (sin contar la declaración DOCTYPE) .

El documento resultante sería el de la Figura 3.40.

```
<!DOCTYPE movies SYSTEM "dtd_modificada.dtd">
<movies>
<film fid="H1">
 <t>Always Tell Your Wife
 <year>1922
 <dirs>
   <dir dirk="R">
     <dirn>Se.Hicks
   <dir dirk="R">
     <dirn>Hitchcock
 <prods>
   <prod prodk="R">
     <pname>Lasky
 <studios>
   <studio>Famous
 <prcs>
   <prc>sbw
 <cats>
   <cat>Dram
<film fid="H2">
 <t>Number Thirteen
 <year>1922
 <dirs>
   <dir dirk="R">
     <dirn>Hitchcock
 <prods>
   <prod prodk="R">
     <pname>Hitchcock
 <studios>
   <studio>Islington
   <distributor>Famous
 <prcs>
   <prc>sbw
   <prctext>unfinished
<film fid="H3">
  <t>Woman to Woman
 <year>1922
 <dirs>
   <dir dirk="R">
     <dirn>Hitchcock
   <dir dirk="N">
     <dirn>Cutts
 <prods>
   <prod prodk="R">
     <pname>Balcon
 <studios>
   <studio>B-S-F
```

```
    <distributor>Wardour
  <prcs>
    <prc>sbw
  <cats>
    <cat>Dram
  <loc>
  <site>
      <siteplace>GB
  <error>same(GCt27), y(1926)
<film fid="H4">
   <t>The Passionate Adventure
  <year>1924
  <dirs>
    <dir dirk="R">
      <dirn>Hitchcock
    <dir dirk="N">
      <dirn>Cutts
  <prods>
    <prod prodk="R">
      <pname>Balcon
  <studios>
    <studio>Gainsborough
    <distributor>GaumontD
  <prcs>
    <prc>sbw
<film fid="H5">
  <t> The Blackguard
  <year>1925
  <dirs>
    <dir dirk="R">
      <dirn>Hitchcock
    <dir dirk="N">
      <dirn>Cutts
  <prods>
    <prod prodk="R">
      <pname>Balcon
  <studios>
    <studio>UFA
    <distributor>Wardour
  <prcs>
    <prc>sbw
```

Figura 3.40: Contenido del archivo registro5_min_omit1.sgml (omisión de etiquetas de fin).

3.3.3.2. Omisión de las etiquetas de inicio

A la hora de omitir una etiqueta de inicio hay que considerar tres condiciones:

1. El elemento debe ser requerido por el contexto y el resto de elementos que puedan aparecer deben ser opcionales.
2. El elemento no puede tener atributos que deba especificar.
3. El elemento no puede tener contenido nulo.

En el documento que se está considerando han podido omitirse bastantes menos etiquetas de inicio que etiquetas de fin debido a que hay muchos elementos opcionales. Para saber si un elemento es requerido por el contexto hay que fijarse de nuevo en la DTD y, más concretamente, en los modelos de contenido de los elementos.

Según la DTD, el primer elemento que debe aparecer es el elemento `movies`, ningún otro puede aparecer en su lugar y, por tanto, su etiqueta de inicio puede omitirse. El contenido del elemento `movies` está formado por elementos de tipo `film`, que podría parecer otro elemento requerido, pero no lo es, pues está afectado por el indicador de ocurrencia '*'. Dicho carácter significa que el elemento puede ocurrir cualquier número de veces, incluso cero. De cualquier manera, tampoco podría omitirse porque la etiqueta de inicio del elemento contiene un atributo que hay que especificar.

El modelo de contenido del elemento `film` está formado por un grupo que usa como conector el carácter ',', eso significa que todos los elementos deben aparecer (a no ser que estén afectados por un indicador de ocurrencia) y deben hacerlo en el orden indicado. Con esto se concluye que, de los subelementos de film, son requeridos por el contexto los elementos t, year y dirs. Las etiquetas de inicio de estos tres elementos pueden ser omitidas y nunca podrán omitirse las etiquetas de inicio del resto de elementos pertenecientes al modelo de contenido de film, aunque sí podrían omitirse las etiquetas de inicio de otros subelementos de los mismos.

El paso siguiente sería estudiar los modelos de contenido de los subelementos de films de la misma forma. Por ejemplo, la etiqueta de inicio del elemento alts no ha podido omitirse, pero sí podrá omitirse la etiqueta de inicio de su subelemento alt, pues está afectado por el indicador de ocurrencia '+', lo que significa que debe aparecer al menos una vez. Eso sí, debe aparecer una vez, pero si aparece más de una vez, las siguientes apariciones no serán requeridas por el contexto, por tanto, sólo puede omitirse la etiqueta de inicio de su primera aparición.

Es un estudio más laborioso, pero no complicado, al estar la DTD formada por elementos cuyos modelos de contenido son similares (grupos de elementos unidos por el conector ',') hay que fijarse en los elementos que no tengan indicador de ocurrencia, cuyas etiquetas de inicio pueden omitirse si no tienen atributos, y en los elementos cuyo indicador de ocurrencia sea '+', cuyas etiquetas de inicio pueden omitirse sólo en su primera aparición dentro del elemento que los contiene.

Resumiendo, pueden omitirse las etiquetas de inicio de los elementos movies, t, year, dirs, altt, dirn, pname, awtype, refdest y refto y, por otro lado, pueden omitirse las etiquetas

de inicio de los elementos alt, studio, prc, cat, aw, site, names y paw pero sólo si es su primera aparición.

El documento resultante es el de la Figura 3.41.

```
<!DOCTYPE movies SYSTEM "dtd_modificada.dtd">
<film fid="H1">
 Always Tell Your Wife</t>
 1922</year>
   <dir dirk="R">
     Se.Hicks</dirn>
   </dir>
   <dir dirk="R">
     Hitchcock</dirn>
   </dir>
 </dirs>
 <prods>
   <prod prodk="R">
     Lasky</pname>
   </prod>
 </prods>
 <studios>
   Famous</studio>
 </studios>
 <prcs>
   sbw</prc>
 </prcs>
 <cats>
   Dram</cat>
 </cats>
</film>
<film fid="H2">
 Number Thirteen</t>
 1922</year>
  <dir dirk="R">
     Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
   <prod prodk="R">
     Hitchcock</pname>
   </prod>
 </prods>
 <studios>
   Islington</studio>
   <distributor>Famous</distributor>
 </studios>
 <prcs>
```

```
    sbw</prc>
    <prctext>unfinished</prctext>
  </prcs>
</film>
<film fid="H3">
 Woman to Woman</t>
 1922</year>
    <dir dirk="R">
      Hitchcock</dirn>
    </dir>
    <dir dirk="N">
      Cutts</dirn>
    </dir>
 </dirs>
 <prods>
    <prod prodk="R">
      Balcon</pname>
    </prod>
 </prods>
 <studios>
    B-S-F</studio>
    <distributor>Wardour</distributor>
 </studios>
 <prcs>
    sbw</prc>
 </prcs>
 <cats>
    Dram</cat>
 </cats>
 <loc>
     <siteplace>GB</siteplace>
    </site>
 </loc>
 <error>same(GCt27), y(1926)</error>
</film>
<film fid="H4">
 The Passionate Adventure</t>
 1924</year>
    <dir dirk="R">
     Hitchcock</dirn>
   </dir>
    <dir dirk="N">
      Cutts</dirn>
    </dir>
 </dirs>
 <prods>
    <prod prodk="R">
```

```
      Balcon</pname>
    </prod>
  </prods>
  <studios>
    Gainsborough</studio>
    <distributor>GaumontD</distributor>
  </studios>
  <prcs>
    sbw</prc>
  </prcs>
</film>
<film fid="H5">
 The Blackguard</t>
 1925</year>
  <dir dirk="R">
     Hitchcock</dirn>
  </dir>
  <dir dirk="N">
     Cutts</dirn>
  </dir>
 </dirs>
 <prods>
   <prod prodk="R">
     Balcon</pname>
   </prod>
 </prods>
 <studios>
   UFA</studio>
   <distributor>Wardour</distributor>
 </studios>
 <prcs>
   sbw</prc>
 </prcs>
</film>
</movies>
```

Figura 3.41: Contenido del archivo registro5_min_omit2.sgml (omisión de etiquetas de inicio).

3.3.3.3. Omisión de etiquetas de inicio y de fin

Para usar conjuntamente la omisión de etiquetas de inicio y de fin no basta con unir las minimizaciones de las dos secciones anteriores. Hay que tener en cuenta los problemas vistos en 3.2.1, y que, en este caso, aparecen, por ejemplo, con el elemento year. Es un elemento requerido por el contexto, con lo cual, en principio, se podría omitir su etiqueta

de inicio, pero si se ha omitido la etiqueta de fin del elemento anterior, el elemento t, el analizador no tendría forma de distinguir los dos elementos, que están formados por datos de usuario y sin ninguna etiqueta entre ellos. Teniendo en cuenta esta circunstancia el resultado es el siguiente (Figura 3.42).

```
<!DOCTYPE movies SYSTEM "dtd_modificada.dtd">
<film fid="H1">Always Tell Your Wife
<year>1922
<dir dirk="R">Se.Hicks
<dir dirk="R"><dirn>Hitchcock
<prods><prod prodk="R">Lasky
<studios>Famous
<prcs>sbw
<cats>Dram
<film fid="H2">Number Thirteen
<year>1922
<dir dirk="R">Hitchcock
<prods><prod prodk="R">Hitchcock
<studios>Islington
<distributor>Famous
<prcs>sbw
<prctext>unfinished
<film fid="H3">Woman to Woman
<year>1922
<dir dirk="R">Hitchcock
<dir dirk="N">Cutts
<prods><prod prodk="R">Balcon
<studios>B-S-F
<distributor>Wardour
<prcs>sbw
<cats>Dram
<loc><siteplace>GB
<error>same(GCt27), y(1926)
<film fid="H4">The Passionate Adventure
<year>1924
<dir dirk="R">Hitchcock
<dir dirk="N">Cutts
<prods><prod prodk="R">Balcon
<studios>Gainsborough
<distributor>GaumontD
<prcs>sbw
<film fid="H5">The Blackguard
<year>1925
<dir dirk="R">Hitchcock
<dir dirk="N">Cutts
<prods><prod prodk="R">Balcon
<studios>UFA
```

```
<distributor>Wardour
<prcs>sbw
</movies>
```

Figura 3.42: Contenido del archivo registro5_min_omit3.sgml (omisión de etiquetas de inicio y de fin).

3.3.4. Aplicación de SHORTTAG

La característica SHORTTAG permite reducir el tamaño de una etiqueta de distintas formas, por ejemplo, permite omitir el separador final algunas etiquetas o las etiquetas de fin de elemento (usando el separador net), pero también hay casos en los que pueden omitirse los identificadores de elemento y sus atributos (o parte de ellos), dejando únicamente los delimitadores. Aquí se hará uso de la mayoría de estas posibilidades, primero una por una, para ver la potencia de cada una de ellas, y luego conjuntamente. Las forma de aplicarlas es la vista en el apartado 3.2.2.

3.3.4.1. Omisión del carácter separador TAGC

La omisión del separador final ('>') de algunas etiquetas es un tipo de minimización con la que no parece que vayan a obtenerse grandes resultados, en cuanto al tanto por ciento de reducción de los caracteres de un archivo (como mucho, un carácter por etiqueta), sin embargo, la técnica para llevarlo a cabo es sencilla y mecánica, podría usarse un programa que elimine estos caracteres, y se obtendría una reducción del tamaño de archivo pequeña, aunque no tan pequeña si se habla de documentos extensos en los que las etiquetas representan un porcentaje elevado del tamaño del archivo. El principal inconveniente a la hora de usar este tipo de minimización puede ser el aumento de la complejidad para leer el documento. Esta opción fue desaprobada por la comunidad SGML[1]. Aún así, se muestra aquí el archivo minimizado, para poder comparar su reducción con el resto de minimizaciones.

El carácter TAGC de una etiqueta sólo podrá omitirse si la etiqueta aparece inmediatamente seguida por otra etiqueta. En la Figura 3.43 puede verse el archivo registro5_min_short1.sgml en el que se han minimizado los primeros cinco registros usando esta posibilidad.

```
<!DOCTYPE movies SYSTEM "dtd_modificada.dtd">
<movies
<film fid="H1"
 <t>Always Tell Your Wife</t
 <year>1922</year
 <dirs
   <dir dirk="R"
     <dirn>Se.Hicks</dirn
```

```
    </dir
    <dir dirk="R"
      <dirn>Hitchcock</dirn
    </dir
  </dirs
  <prods
    <prod prodk="R"
      <pname>Lasky</pname
    </prod
  </prods
  <studios
    <studio>Famous</studio
  </studios
  <prcs
    <prc>sbw</prc
  </prcs
  <cats
    <cat>Dram</cat
  </cats
</film
<film fid="H2"
  <t>Number Thirteen</t
  <year>1922</year
  <dirs
    <dir dirk="R"
      <dirn>Hitchcock</dirn
    </dir
  </dirs
  <prods
    <prod prodk="R"
      <pname>Hitchcock</pname
    </prod
  </prods
  <studios
    <studio>Islington</studio
    <distributor>Famous</distributor
  </studios
  <prcs
    <prc>sbw</prc
    <prctext>unfinished</prctext
  </prcs
</film
<film fid="H3"
  <t>Woman to Woman</t
  <year>1922</year
  <dirs
    <dir dirk="R"
```

```
      <dirn>Hitchcock</dirn
    </dir
    <dir dirk="N"
      <dirn>Cutts</dirn
    </dir
  </dirs
  <prods
    <prod prodk="R"
      <pname>Balcon</pname
    </prod
  </prods
  <studios
    <studio>B-S-F</studio
    <distributor>Wardour</distributor
  </studios
  <prcs
    <prc>sbw</prc
  </prcs
  <cats
    <cat>Dram</cat
  </cats
  <loc
    <site
      <siteplace>GB</siteplace
    </site
  </loc
  <error>same(GCt27), y(1926)</error
</film
<film fid="H4"
  <t>The Passionate Adventure</t
  <year>1924</year
  <dirs
    <dir dirk="R"
      <dirn>Hitchcock</dirn
  </dir
    <dir dirk="N"
      <dirn>Cutts</dirn
    </dir
  </dirs
  <prods
    <prod prodk="R"
      <pname>Balcon</pname
    </prod
  </prods
  <studios
    <studio>Gainsborough</studio
    <distributor>GaumontD</distributor
```

```
 </studios
 <prcs
   <prc>sbw</prc
 </prcs
</film
<film fid="H5"
 <t>The Blackguard</t
 <year>1925</year
 <dirs
   <dir dirk="R"
     <dirn>Hitchcock</dirn
 </dir
   <dir dirk="N"
     <dirn>Cutts</dirn
   </dir
 </dirs
 <prods
   <prod prodk="R"
     <pname>Balcon</pname
   </prod
 </prods
 <studios
   <studio>UFA</studio
   <distributor>Wardour</distributor
 </studios
 <prcs
   <prc>sbw</prc
 </prcs
</film
</movies>
```

Figura 3.43: Contenido del archivo registro5_min_short1.sgml (omisión del carácter TAGC).

3.3.4.2. Omisión del identificador de la etiqueta

La siguiente opción es la de omitir los identificadores de las etiquetas de fin cuando sea posible y, será posible, cuando cada uno pueda completarse con el identificador de la etiqueta de inicio más cercana. Es evidente que la reducción sería aún mayor si las etiquetas de fin se eliminasen por completo, como puede hacerse con omittag, sin embargo, esta opción puede ser más sencilla tanto para usuarios como para desarrolladores y también supone una buena reducción del documento, por lo que debe tenerse en cuenta, de hecho, es una de las características cuya inclusión fue motivo de debate durante el desarrollo del lenguaje XML.

Los autores de XML coinciden en que la dificultad no está en el parser sino en la posible complejidad para el usuario [46]. Completar cada etiqueta de fin vacía con la etiqueta de inicio más cercana no es una tarea muy complicada para un analizador, podría hacerse con una pila que almacenase los identificadores y así controlar qué elemento debe cerrarse. El principal argumento en contra de esta opción es que complica la lectura del documento, sin embargo, en casos como este, en los que se habla de una base de datos, no resulta menos legible.

Una de las opciones propuestas para mantener el uso de esta opción en XML fue la de usarla solamente en los nodos hoja, aquellos elementos que no tienen más subelementos, ya que por la proximidad entre la etiqueta de inicio y la de fin no afectaría, en absoluto, a la lectura del documento, sin embargo, en aquel momento se consideró que las ventajas que podían obtenerse en las bases de datos no eran suficientes, a pesar de la gran importancia que tiene recuperar información de bases de datos en la Web. Además de la pequeña complejidad añadida para el analizador, se pensó que dificultaría la detección de errores y se consideró la "redundancia" una ventaja para disminuir errores en las transmisiones de datos, aunque realmente no sea ese su objetivo.

Al aplicar esta característica de minimización se han podido omitir todos los identificadores de las etiquetas finales. Las etiquetas de fin siempre pueden completarse con las etiquetas de inicio más cercanas. En la teoría se ha visto que también pueden omitirse los identificadores de las etiquetas de inicio, sin embargo, esta posibilidad no puede aplicarse a ninguno de estos cinco registros puesto que, al trabajar con omittag habilitado, la etiqueta debería completarse con el identificador de la etiqueta de inicio más cercana. Esto sólo tiene utilidad si hay varios elementos iguales seguidos, cosa que no ocurre en este ejemplo. Sí se hará uso de esta posibilidad cuando se trabaje con los cincuenta registros completos. En la Figura 3.44 puede verse el archivo registro5_min_short2.sgml en el que se han minimizado los primeros cinco registros omitiendo los identificadores de las etiquetas de fin.

```
<!DOCTYPE movies SYSTEM "dtd_modificada.dtd">
<movies>
<film fid="H1">
 <t>Always Tell Your Wife</>
 <year>1922</>
 <dirs>
   <dir dirk="R">
     <dirn>Se.Hicks</>
   </>
   <dir dirk="R">
     <dirn>Hitchcock</>
   </>
 </>
 <prods>
   <prod prodk="R">
     <pname>Lasky</>
   </>
 </>
```

```
 <studios>
   <studio>Famous</>
 </>
 <prcs>
   <prc>sbw</>
 </>
 <cats>
   <cat>Dram</>
 </>
</>
<film fid="H2">
 <t>Number Thirteen</>
 <year>1922</>
 <dirs>
   <dir dirk="R">
     <dirn>Hitchcock</>
   </>
 </>
 <prods>
   <prod prodk="R">
     <pname>Hitchcock</>
   </>
 </>
 <studios>
   <studio>Islington</>
   <distributor>Famous</>
 </>
 <prcs>
   <prc>sbw</>
   <prctext>unfinished</>
 </>
</>
<film fid="H3">
 <t>Woman to Woman</>
 <year>1922</>
 <dirs>
   <dir dirk="R">
     <dirn>Hitchcock</>
   </>
   <dir dirk="N">
     <dirn>Cutts</>
   </>
 </>
 <prods>
   <prod prodk="R">
     <pname>Balcon</>
   </>
```

```
 </>
 <studios>
   <studio>B-S-F</>
   <distributor>Wardour</>
 </>
 <prcs>
   <prc>sbw</>
 </>
 <cats>
   <cat>Dram</>
 </>
 <loc>
   <site>
     <siteplace>GB</>
   </>
 </>
 <error>same(GCt27), y(1926)</>
</>
<film fid="H4">
 <t>The Passionate Adventure</>
 <year>1924</>
 <dirs>
   <dir dirk="R">
     <dirn>Hitchcock</>
 </>
   <dir dirk="N">
     <dirn>Cutts</>
   </>
 </>
 <prods>
   <prod prodk="R">
     <pname>Balcon</>
   </>
 </>
 <studios>
   <studio>Gainsborough</>
   <distributor>GaumontD</>
 </>
 <prcs>
   <prc>sbw</>
 </>
</>
<film fid="H5">
 <t>The Blackguard</>
 <year>1925</>
 <dirs>
   <dir dirk="R">
```

```
    <dirn>Hitchcock</>
 </>
   <dir dirk="N">
    <dirn>Cutts</>
   </>
 </>
 <prods>
   <prod prodk="R">
    <pname>Balcon</>
   </>
 </>
 <studios>
   <studio>UFA</>
   <distributor>Wardour</>
 </>
 <prcs>
   <prc>sbw</>
 </>
</>
</>
```

Figura 3.44: Contenido del archivo registro5_min_short2.sgml (omisión del identificador en las etiquetas de fin).

3.3.4.3. net

Otro tipo de minimización, con la que se obtienen mejores resultados que con la anterior, consiste en sustituir la etiqueta de fin completa por el separador "null end tag", que suele ser el carácter "/". En ese caso sería necesario sustituir también el separador final de la etiqueta de inicio correspondiente por dicho carácter. Aquí sí se aprecia una clara desventaja en la dificultad para leer el documento. En la Figura 3.45 puede verse el archivo registro5_min_short3.sgml en el que se han minimizado los primeros cinco registros usando esta posibilidad.

```
<!DOCTYPE movies SYSTEM "dtd_modificada.dtd">
<movies/
<film fid="H1"/
 <t/Always Tell Your Wife/
 <year/1922/
 <dirs/
   <dir dirk="R"/
    <dirn/Se.Hicks/
   /
   <dir dirk="R"/
```

```
      <dirn/Hitchcock/
    /
  /
  <prods/
    <prod prodk="R"/
      <pname/Lasky/
    /
  /
  <studios/
    <studio/Famous/
  /
  <prcs/
    <prc/sbw/
  /
  <cats/
    <cat/Dram/
  /
/
<film fid="H2"/
 <t/Number Thirteen/
 <year/1922/
 <dirs/
   <dir dirk="R"/
     <dirn/Hitchcock/
   /
 /
 <prods/
   <prod prodk="R"/
     <pname/Hitchcock/
   /
 /
 <studios/
   <studio/Islington/
   <distributor/Famous/
 /
 <prcs/
   <prc/sbw/
   <prctext/unfinished/
 /
/
<film fid="H3"/
 <t/Woman to Woman/
 <year/1922/
 <dirs/
   <dir dirk="R"/
     <dirn/Hitchcock/
   /
```

```
    <dir dirk="N"/
      <dirn/Cutts/
    /
  /
  <prods/
    <prod prodk="R"/
      <pname/Balcon/
    /
  /
  <studios/
    <studio/B-S-F/
    <distributor/Wardour/
/
  <prcs/
    <prc/sbw/
  /
  <cats/
    <cat/Dram/
  /
  <loc/
    <site/
      <siteplace/GB/
    /
  /
  <error/same(GCt27), y(1926)/
/
<film fid="H4"/
  <t/The Passionate Adventure/
  <year/1924/
  <dirs/
    <dir dirk="R"/
      <dirn/Hitchcock/
  /
    <dir dirk="N"/
      <dirn/Cutts/
    /
  /
  <prods/
    <prod prodk="R"/
      <pname/Balcon/
    /
  /
  <studios/
    <studio/Gainsborough/
    <distributor/GaumontD/
  /
  <prcs/
```

```
    <prc/sbw/
  /
  /
  <film fid="H5"/
   <t/The Blackguard/
   <year/1925/
   <dirs/
     <dir dirk="R"/
       <dirn/Hitchcock/
   /
     <dir dirk="N"/
       <dirn/Cutts/
    /
   /
   <prods/
     <prod prodk="R"/
       <pname/Balcon/
     /
   /
   <studios/
     <studio/UFA/
     <distributor/Wardour/
  /
   <prcs/
     <prc/sbw/
   /
  /
  /
```

Figura 3.45: Contenido del archivo registro5_min_short3.sgml (uso del separador net).

3.3.4.4. Omisión de la lista de atributos

La siguiente posibilidad que se tiene en cuenta es la de minimizar la lista de atributos. El ejemplo que se está considerando no incluye muchos atributos, con lo que minimizarlos aquí no supone una gran variación del tamaño de archivo, pero sí puede serlo en otras bases de datos donde se haga más uso de los mismos.

Lo más sencillo es optar por eliminar solamente las comillas que delimitan el valor de cada atributo (si dicho valor está compuesto solo por name characters), pero, como se vio en la sección 3.2.2.4, algunos atributos podrían quedar reducidos a su valor o incluso omitirse por completo. En la Figura 3.46 se muestra el contenido del archivo registro5_min_short4.sgml donde se omiten sólo las comillas de los atributos.

```
<!DOCTYPE movies SYSTEM "dtd_modificada.dtd">
```

```
<movies>
<film fid=H1>
 <t>Always Tell Your Wife</t>
 <year>1922</year>
 <dirs>
   <dir dirk=R>
     <dirn>Se.Hicks</dirn>
   </dir>
   <dir dirk=R>
     <dirn>Hitchcock</dirn>
   </dir>
 </dirs>
 <prods>
   <prod prodk=R>
     <pname>Lasky</pname>
   </prod>
 </prods>
 <studios>
   <studio>Famous</studio>
 </studios>
 <prcs>
   <prc>sbw</prc>
 </prcs>
 <cats>
   <cat>Dram</cat>
 </cats>
</film>
<film fid=H2>
 <t>Number Thirteen</t>
 <year>1922</year>
 <dirs>
   <dir dirk=R>
     <dirn>Hitchcock</dirn>
   </dir>
 </dirs>
 <prods>
   <prod prodk=R>
     <pname>Hitchcock</pname>
   </prod>
 </prods>
 <studios>
   <studio>Islington</studio>
   <distributor>Famous</distributor>
 </studios>
 <prcs>
   <prc>sbw</prc>
   <prctext>unfinished</prctext>
```

```
 </prcs>
</film>
<film fid=H3>
 <t>Woman to Woman</t>
 <year>1922</year>
 <dirs>
   <dir dirk=R>
     <dirn>Hitchcock</dirn>
   </dir>
   <dir dirk=N>
     <dirn>Cutts</dirn>
   </dir>
 </dirs>
 <prods>
   <prod prodk=R>
     <pname>Balcon</pname>
   </prod>
 </prods>
 <studios>
   <studio>B-S-F</studio>
   <distributor>Wardour</distributor>
 </studios>
 <prcs>
   <prc>sbw</prc>
 </prcs>
 <cats>
   <cat>Dram</cat>
 </cats>
 <loc>
   <site>
     <siteplace>GB</siteplace>
   </site>
 </loc>
 <error>same(GCt27), y(1926)</error>
</film>
<film fid=H4>
 <t>The Passionate Adventure</t>
 <year>1924</year>
 <dirs>
   <dir dirk=R>
     <dirn>Hitchcock</dirn>
 </dir>
   <dir dirk=N>
     <dirn>Cutts</dirn>
   </dir>
 </dirs>
 <prods>
```

```
    <prod prodk=R>
      <pname>Balcon</pname>
    </prod>
  </prods>
  <studios>
    <studio>Gainsborough</studio>
    <distributor>GaumontD</distributor>
  </studios>
  <prcs>
    <prc>sbw</prc>
  </prcs>
</film>
<film fid=H5>
  <t>The Blackguard</t>
  <year>1925</year>
  <dirs>
    <dir dirk=R>
      <dirn>Hitchcock</dirn>
  </dir>
    <dir dirk=N>
      <dirn>Cutts</dirn>
    </dir>
  </dirs>
  <prods>
    <prod prodk=R>
      <pname>Balcon</pname>
    </prod>
  </prods>
  <studios>
    <studio>UFA</studio>
    <distributor>Wardour</distributor>
  </studios>
  <prcs>
    <prc>sbw</prc>
  </prcs>
</film>
</movies>
```

Figura 3.46: Contenido del archivo registro5_min_short4.sgml (omisión de las comillas en los valores de atributos).

En la Figura 3.47 se muestra un archivo en el que se deja sólo el valor del atributo para aquellos casos en que tengan declarada una lista de posibles valores. Si el valor coincide con el valor por defecto, el atributo podría eliminarse por completo. No se han eliminado las comillas de los atributos que no se encuentran en esta situación por seguir viendo cada tipo de minimización por separado.

```
<!DOCTYPE movies SYSTEM "dtd_modificada.dtd">
<movies>
<film fid="H1">
 <t>Always Tell Your Wife</t>
 <year>1922</year>
 <dirs>
   <dir R>
     <dirn>Se.Hicks</dirn>
   </dir>
   <dir R>
     <dirn>Hitchcock</dirn>
   </dir>
 </dirs>
 <prods>
   <prod R>
     <pname>Lasky</pname>
   </prod>
 </prods>
 <studios>
   <studio>Famous</studio>
 </studios>
 <prcs>
   <prc>sbw</prc>
 </prcs>
 <cats>
   <cat>Dram</cat>
 </cats>
</film>
<film fid="H2">
 <t>Number Thirteen</t>
 <year>1922</year>
 <dirs>
   <dir R>
     <dirn>Hitchcock</dirn>
   </dir>
 </dirs>
 <prods>
   <prod R>
     <pname>Hitchcock</pname>
   </prod>
 </prods>
 <studios>
   <studio>Islington</studio>
   <distributor>Famous</distributor>
 </studios>
 <prcs>
   <prc>sbw</prc>
```

```
    <prctext>unfinished</prctext>
  </prcs>
 </film>
 <film fid="H3">
  <t>Woman to Woman</t>
  <year>1922</year>
  <dirs>
    <dir R>
      <dirn>Hitchcock</dirn>
    </dir>
    <dir>
      <dirn>Cutts</dirn>
    </dir>
  </dirs>
  <prods>
    <prod R>
      <pname>Balcon</pname>
    </prod>
  </prods>
  <studios>
    <studio>B-S-F</studio>
    <distributor>Wardour</distributor>
  </studios>
  <prcs>
    <prc>sbw</prc>
  </prcs>
  <cats>
    <cat>Dram</cat>
  </cats>
  <loc>
    <site>
      <siteplace>GB</siteplace>
    </site>
  </loc>
  <error>same(GCt27), y(1926)</error>
 </film>
 <film fid="H4">
  <t>The Passionate Adventure</t>
  <year>1924</year>
  <dirs>
    <dir R>
      <dirn>Hitchcock</dirn>
  </dir>
    <dir>
      <dirn>Cutts</dirn>
    </dir>
  </dirs>
```

```
 <prods>
   <prod R>
     <pname>Balcon</pname>
   </prod>
 </prods>
 <studios>
   <studio>Gainsborough</studio>
   <distributor>GaumontD</distributor>
 </studios>
 <prcs>
   <prc>sbw</prc>
 </prcs>
</film>
<film fid="H5">
 <t>The Blackguard</t>
 <year>1925</year>
 <dirs>
   <dir R>
     <dirn>Hitchcock</dirn>
 </dir>
   <dir>
     <dirn>Cutts</dirn>
   </dir>
 </dirs>
 <prods>
   <prod R>
     <pname>Balcon</pname>
   </prod>
 </prods>
 <studios>
   <studio>UFA</studio>
   <distributor>Wardour</distributor>
 </studios>
 <prcs>
   <prc>sbw</prc>
 </prcs>
</film>
</movies>
```

Figura 3.47: Contenido del archivo registro5_min_short5.sgml (omisión de parte de la lista de atributos).

3.3.4.5. Aplicación de todas las posibilidades de shorttag

En la Figura 3.48 se verá el resultado de aplicar las posibilidades de minimización que ofrece shorttag. No se intenta usar todas las posibles, sino combinarlas de forma que se

obtengan los mejores resultados. Para ello, se ha considerado preferible empezar usando net, que sustituye toda la etiqueta de fin por un carácter y, por tanto, es más efectivo que omitir el identificador, si sólo se tiene en cuenta la reducción de tamaño. Eso imposibilita usar la omisión de tagc. Posteriormente se aplica la minimización de los atributos de la forma vista en las secciones anteriores.

```
<!DOCTYPE movies SYSTEM "dtd_modificada.dtd">
<movies/
<film fid=H1/
 <t/Always Tell Your Wife/
 <year/1922/
 <dirs/
   <dir R/
     <dirn/Se.Hicks/
   /
   <dir R/
     <dirn/Hitchcock/
   /
 /
 <prods/
   <prod R/
     <pname/Lasky/
   /
 /
 <studios/
   <studio/Famous/
 /
 <prcs/
   <prc/sbw/
 /
 <cats/
   <cat/Dram/
 /
/
<film fid=H2/
 <t/Number Thirteen/
 <year/1922/
 <dirs/
   <dir R/
     <dirn/Hitchcock/
   /
 /
 <prods/
   <prod R/
     <pname/Hitchcock/
   /
 /
```

```
 <studios/
   <studio/Islington/
   <distributor/Famous/
 /
 <prcs/
   <prc/sbw/
   <prctext/unfinished/
 /
/
<film fid=H3/
 <t/Woman to Woman/
 <year/1922/
 <dirs/
   <dir R/
     <dirn/Hitchcock/
   /
   <dir/
     <dirn/Cutts/
   /
 /
 <prods/
   <prod R/
     <pname/Balcon/
   /
 /
 <studios/
   <studio/B-S-F/
   <distributor/Wardour/
/
 <prcs/
   <prc/sbw/
 /
 <cats/
   <cat/Dram/
 /
 <loc/
   <site/
     <siteplace/GB/
   /
 /
 <error/same(GCt27), y(1926)/
/
<film fid=H4/
 <t/The Passionate Adventure/
 <year/1924/
 <dirs/
   <dir R/
```

```
      <dirn/Hitchcock/
  /
    <dir/
      <dirn/Cutts/
    /
  /
  <prods/
    <prod R/
      <pname/Balcon/
    /
  /
  <studios/
    <studio/Gainsborough/
    <distributor/GaumontD/
  /
  <prcs/
    <prc/sbw/
/
/
<film fid=H5/
 <t/The Blackguard/
 <year/1925/
 <dirs/
   <dir R/
     <dirn/Hitchcock/
 /
   <dir/
     <dirn/Cutts/
  /
 /
 <prods/
   <prod R/
     <pname/Balcon/
   /
 /
 <studios/
   <studio/UFA/
   <distributor/Wardour/
/
 <prcs/
   <prc/sbw/
 /
/
/
```

Figura 3.48: Contenido del archivo registro5_min_short6.sgml (uso de todas las minimizaciones permitidas por la característica Shorttag).

3.3.5. SHORTREF, DATATAG y RANK

Estas tres opciones de SGML no se han aplicado a los archivos con los que se está trabajando. El motivo de no haber usado la característica RANK es claro, se trata una característica que permite omitir el nivel de anidamiento de algunas etiquetas y, en este caso, no hay ningún identificador de elemento que especifique su nivel de anidamiento.

En cuanto a la característica SHORTREF, es una de las opciones con las que más se podría reducir el marcado en ciertas bases de datos. Esta opción podría implementarse de forma sencilla en bases de datos cuya estructura fuese la de una tabla con un número fijo de columnas. Sin embargo, la base de datos con la que se trabaja está repleta de elementos opcionales, que hacen poco práctico el uso de SHORTREF. En las Figuras 3.49 y 3.50 se ha incluido un ejemplo para ilustrarlo mejor.

```
<!ENTITY  filmt    "<film><t>">
<!ENTITY  tyear    "</t><year>">
<!ENTITY  yeardirn "</year><dirn>">
<!ENTITY  dirnfilm "</dirn></film>">

<!SHORTREF mapa1   "(" filmt
                   ")" dirnfilm >
<!SHORTREF mapa2   "," tyear>
<!SHORTREF mapa3   "," yeardirn>

<!USEMAP   mapa1    movies>
<!USEMAP   mapa2    t>
<!USEMAP   mapa3    year>

<!ELEMENT  movies o o  (film*)>
<!ELEMENT  film   o o  (t, year, dirn)>
<!ELEMENT  t      o o  (#PCDATA)>
<!ELEMENT  year   o o  (#PCDATA)>
<!ELEMENT  dirn   o o  (#PCDATA)>
```

Figura 3.49: Contenido del archivo dtd_shortref.dtd.

```
<!DOCTYPE movies SYSTEM "dtd_shortref.dtd">
<movies>
  (Always Tell Your Wife, 1922, Se.Hicks)
```

```
  (Number Thirteen, 1922, Hitchcock)
  (Woman to Woman, 1922, Hitchcock)
  (The Passionate Adventure, 1924, Hitchcock)
  (The Blackguard, 1925, Hitchcock)
</movies>
```

Figura 3.50: Contenido del archivo registro5_shortref.sgml (uso de shortref).

La Figura 3.50 muestra lo que podrían ser cinco registros de una pequeña base de datos que almacenase información sobre título, fecha y director de varias películas. El marcado ha quedado reducido a los caracteres "(" ,")" y ",", gracias al uso que se hace de SHORTREF en la DTD de la Figura 3.49. El carácter "," sustituye a distintas etiquetas según el lugar donde aparezca, debido a los distintos mapas de referencias que se activan según el elemento que se considere.

Hasta aquí sería sencillo, sin embargo, en el ejemplo con el que se trabaja en este proyecto, muchos elementos tienen subelementos, que a su vez tienen más subelementos. Esto significa que no sería suficiente con usar el carácter "," para separarlos, habría que usar diferentes caracteres separadores en los distintos niveles de anidamiento para poder diferenciarlos, con lo que la lectura del documento se volvería más compleja. Además, el documento con el que se trabaja tiene muchos elementos opcionales lo que complicaría mucho más el trabajo. Si después de un elemento no se tiene la certeza de cuál aparecerá a continuación, serían necesarios más caracteres para diferenciarlos, con lo que la característica SHORTREF deja de ser práctica.

El problema sería similar con el uso de la característica DATATAG. Es muy útil para datos que siguen un patrón fijo, pero no en este caso. Además, como se ha visto en 3.1, esta característica no fue implementada por los analizadores estudiados al considerar que puede ser sustituida completamente por SHORTREF.

3.3.6. Aplicación de OMITTAG y SHORTTAG conjuntamente

En esta sección se aplican todas las posibilidades de minimización disponibles gracias a las características Omittag y Shorttag. El orden en que se han aplicado no es arbitrario. El uso de algunas de ellas puede imposibilitar o hacer innecesario el uso de otras muchas. Se decide empezar omitiendo las etiquetas de fin pues es la opción con la que más se ha reducido el documento. A continuación, se omiten las etiquetas de inicio para los casos vistos en la Sección 3.3.6 y, se buscan entre las restantes, aquellas en las que pueda omitirse, al menos, el identificador. Todavía hay algunas etiquetas que aparecen seguidas, por lo que pueden reducirse algunos caracteres más usando la omisión del carácter tagc. Por último, se reduce, donde es posible, la lista de atributos. Los atributos cuyo valor coincida con el valor por defecto especificado en la DTD se omiten por completo, aquellos que tengan un valor contenido en una lista de valores quedan reducidos a su valor y, para los que no se encuentren en ninguno de los casos anteriores, pueden omitirse las comillas que encierran dicho valor.

Esta es la reducción del documento que se ha considerado óptima y es la que se aplica, en primer lugar al documento con los cinco registros y, a continuación, al documento completo con cincuenta registros.

3.3.6.1. Aplicación a cinco registros

En la Figura 3.51 se aplican todas las minimizaciones contempladas a los cinco primeros registros. Aquí puede apreciarse una disminución significativa en la longitud del documento sin necesidad de hacer ninguna medida.

```
<!DOCTYPE movies SYSTEM "dtd_modificada.dtd">
<film fid=H1>Always Tell Your Wife
 <year>1922
 <dir R>Se.Hicks
 <dir R>Hitchcock
 <prods<prod R>Lasky
 <studios>Famous
 <prcs>sbw
 <cats>Dram
<film fid=H2>Number Thirteen
 <year>1922
 <dir R>Hitchcock
 <prods<prod R>Hitchcock
 <studios>Islington
 <distributor>Famous
 <prcs>sbw
 <prctext>unfinished
<film fid=H3>Woman to Woman
 <year>1922
 <dir R>Hitchcock
 <dir>Cutts
 <prods<prod R>Balcon
 <studios>B-S-F
 <distributor>Wardour
 <prcs>sbw
 <cats>Dram
 <loc<siteplace>GB
 <error>same(GCt27), y(1926)
<film fid=H4>The Passionate Adventure
 <year>1924
 <dir R>Hitchcock
 <dir>Cutts
 <prods<prod R>Balcon
 <studios>Gainsborough
 <distributor>GaumontD
 <prcs>sbw
<film fid=H5>The Blackguard
 <year>1925
 <dir R>Hitchcock
 <dir>Cutts
 <prods<prod R>Balcon
```

```
<studios>UFA
<distributor>Wardour
<prcs>sbw
```

Figura 3.51: Contenido del archivo registro5_min_oys.sgml (uso de todas las minimizaciones permitidas por las características Shorttag y Omittag).

3.3.6.2. Aplicación a cincuenta registros

La aplicación de las minimizaciones a cincuenta registros, a simple vista, también reduce notablemente la longitud del archivo. El resultado se incluye en el Anexo A.2 debido a su extensión. La medida en que se reduce cada uno de los registros no tiene porqué ser la misma, no sólo por la longitud de los datos que contienen, sino porque hay muchos elementos opcionales. Esto hace que la cantidad de elementos contenidos en cada registro pueda ser muy variable. El hecho de que haya registros con distintos tipos de elementos ha permitido aplicar algunas minimizaciones que no se usaron con los primeros cinco registros, por ejemplo, la omisión del identificador de algunas etiquetas de inicio. Será en el siguiente capítulo (Capítulo 4) donde se vean los resultados numéricos de esta reducción de etiquetas.

3.4. Conclusiones

En este capítulo se han estudiado las características de minimización del lenguaje SGML. Cada característica se ha definido y se ha aplicado a un ejemplo sencillo para observar los problemas que pueden surgir en su aplicación. Posteriormente, se han aplicado, una a una, a un archivo extraído de una base de datos. Como resultado, se han generado diez archivos minimizados distintos partiendo de un documento que contiene un registro de la base de datos, otros diez, partiendo de un documento que contiene cinco registros y, finalmente, un archivo minimizado para el documento que incluye cincuenta registros. Para aplicar las minimizaciones a este último documento, se ha buscado la combinación de las mismas que más consigue reducirlo. Una vez generados los archivos, se ha comprobado la validez de todos ellos usando el analizador incluido en OpenSP. Estos archivos generados se utilizarán en el capítulo siguiente para comparar numéricamente las reducciones de tamaño conseguidas en los mismos.

Capítulo 4

Medidas y resultados

En este capítulo se comprobarán numéricamente las reducciones de tamaño conseguidas con la aplicación de las minimizaciones. Contando el número de caracteres de cada archivo antes y después de cada minimización, o, simplemente, viendo la diferencia de tamaño que ocupa en memoria, sería suficiente para conocer el tanto por ciento de reducción de archivo que se consigue con cada tipo de minimización o el número de caracteres de etiquetas que se han eliminado, pero sería interesante conocer qué porcentaje suponen esas etiquetas eliminadas respecto al total de etiquetas que tiene el archivo original y, para ello, es necesario contar los caracteres de etiquetas de dicho archivo. Para facilitar la obtención de resultados se decide escribir un programa en lenguaje Java. El programa leerá los archivos contenidos en un directorio que se le pasa como argumento, asociará cada archivo completo con sus archivos minimizados y devolverá los resultados por pantalla.

En la primera sección del capítulo se describe el programa Java desarrollado y se incluye la salida que genera para los archivos que se desea medir. A continuación, en la sección 4.2, se estudian los resultados obtenidos y se muestran en forma de gráficas para facilitar la comparación de los mismos.

4.1. Programa desarrollado y medidas obtenidas

Se ha desarrollado un programa en lenguaje Java que permite conocer la reducción de caracteres de etiquetas obtenida con los archivos generados en la Sección 3.3. Los archivos minimizados, junto con los originales, se colocan en un directorio que se pasa como argumento al programa y éste devuelve por pantalla el número de caracteres totales y de etiquetas de todos ellos, además del tanto por ciento de reducción de caracteres totales y de etiquetas en los archivos minimizados. El objetivo del programa ha sido facilitar la obtención de estos resultados y, para que los resultados sean correctos, los archivos deben ser archivos SGML válidos (ya han sido analizados con OpenSP) y deben seguir un criterio de nombres, que se ha fijado con el fin de distinguir los archivos minimizados que corresponden a cada archivo completo.

El programa está dividido en dos ficheros que contienen las clases DatosFichero y ComparaCaracteres. La última es la que contiene el método main. Ambas clases se describen en los siguientes apartados.

4.1.1. Clase DatosFichero

La clase DatosFichero contiene dos variables, *ruta* y *minimizado*, que almacenan información sobre la ruta de cada fichero y si éste está, o no, minimizado. Además incluye los métodos cuentaCaracteres y cuentaCaracteresEtiquetas. Ambos métodos utilizan la clase FileReader para leer el archivo y contar caracteres. El primero cuenta los caracteres del archivo descartando caracteres de nueva línea, retornos de carro, tabuladores y espacios en blanco, como puede verse en la Figura 4.1.

```
FileReader f=new FileReader(ruta);
while((c=f.read())!=-1){
      if((char)c!='\n'&& (char)c!='\r' && (char)c!='\t' && (char)c!=' ')
      caracteres++;
}
```

Figura 4.1: Fragmento del método cuentaCaracteres.

El segundo método, cuentaCaracteresEtiquetas, sólo cuenta aquellos caracteres contenidos entre los delimitadores '<' y '>' (incluyendo los propios delimitadores). Esta forma de contar los caracteres de etiquetas no sería válida para algunos de los archivos minimizados, sin embargo, es suficiente con poder contar los caracteres de etiquetas en los archivos completos. En la Sección 4.1.2 se verá la forma de calcular los caracteres de etiquetas de los archivos minimizados.

Se utiliza una sentencia *switch* para determinar las acciones a realizar según el carácter leído (Figura 4.2) y una variable de tipo *boolean*, la variable *etiqueta*, para indicar si el carácter considerado forma parte, o no, de una etiqueta.

Si el carácter leído resulta ser el de inicio de etiqueta ('<') se le da el valor *'true'* a la variable *etiqueta* y se incrementa el valor de la variable que almacena el número de caracteres de la etiqueta actual. Por el contrario, si el carácter que se lee es el de fin de etiqueta ('>'), se suman los caracteres de la etiqueta actual al total de caracteres de etiquetas del archivo, y la variable *etiqueta* pasa a tener el valor *'false'*. Para el resto de caracteres, se comprueba el valor de la variable *etiqueta* con el fin de determinar si se incrementa el contador.

```
switch(c){
        case '<':
          if(!etiqueta){
            etiqueta=true;
            caracteres_etiqueta_actual++;
          }
          break;
      case '>':
          if(etiqueta){
            etiqueta=false;
            caracteres_etiqueta_actual++;
```

```
            caracteres_etiquetas+=caracteres_etiqueta_actual;
            caracteres_etiqueta_actual=0;
          }
          break;
        default:
          if(etiqueta && (char)c!='\n'&& (char)c!='\r' && (char)c!='\t'
            && (char)c!=' ')
            caracteres_etiqueta_actual++;
        break;
}
```

Figura 4.2: Fragmento del método cuentaCaracteresEtiquetas.

En el Anexo B.1 se incluye el código completo correspondiente a esta clase.

4.1.2. Clase ComparaCaracteres

Esta clase contiene el método main. En dicho método se calcularán las reducciones de los archivos con ayuda de los métodos definidos en la clase DatosFichero.

Se comienza comprobando que la entrada conste de un solo argumento y que se trate de un directorio. Después, se examinan los nombres de los archivos contenidos en el directorio, para diferenciar entre archivos completos y archivos minimizados (Figura 4.3). Para poder establecer esta diferencia, se decide que el nombre de los archivos minimizados siga un patrón fijo. Sus nombres deben estar formados por el nombre del archivo completo del que proceden, la cadena "min", que los identifica como minimizados, y cualquier cadena que indique el tipo de minimización, todo ello separado por caracteres "_" y acabado con la extensión ".sgml". Por ejemplo, un archivo completo llamado registro1.sgml puede tener un archivo minimizado llamado registro1_min_omit1.sgml.

```
if(lista_ficheros[i].indexOf("_min_")!=-1){
            // Es un archivo que ha sido minimizado.
            ficheros_minimizados.add(lista_ficheros[i]);

}else{
            // Es un archivo que no ha sido minimizado.
            ficheros_completos.add(lista_ficheros[i]);
}
```

Figura 4.3: Fragmento de código en el que se diferencia entre archivos completos y minimizados.

Una vez diferenciados los archivos, se cuentan los caracteres totales y los caracteres correspondientes a etiquetas de cada archivo completo con el siguiente código:

```
DatosFichero datosfichero1 = new DatosFichero(rutaFicheroCompleto,false);
caracteresCompleto=datosfichero1.cuentaCaracteres();
caracteresEtiquetasCompleto = datosfichero1.cuentaCaracteresEtiquetas();
```

A continuación, se comprueba cuáles de los archivos minimizados provienen del archivo completo que se está considerando, para ello se comparan sus nombres de archivo en la siguiente sentencia if:

```
if(campos_minimizado[0].equals(campos_completo[0]))
```

Posteriormente, se cuentan los caracteres totales de cada archivo minimizado que haya sido obtenido a partir del archivo completo mencionado y se determina su número de caracteres de etiquetas, restando a los caracteres de etiquetas del archivo completo la reducción de caracteres conseguida (los caracteres eliminados al minimizar siempre son caracteres de etiquetas), como puede verse en las siguientes líneas de código:

```
caracteresMinimizado=datosfichero2.cuentaCaracteres();
caracteresEtiquetasMinimizado=caracteresEtiquetasCompleto
      -(caracteresCompleto-caracteresMinimizado);
```

Por último, se calcula el porcentaje de reducción de etiquetas y el de reducción del total de caracteres de archivo (Figura 4.4) y se muestran los resultados por pantalla (Figura 4.5).

```
reduccionArchivo=100.0-((caracteresMinimizado*100.0)
      /caracteresCompleto);
reduccionEtiquetas=100.0-((caracteresEtiquetasMinimizado*100.0)
      /caracteresEtiquetasCompleto);
```

Figura 4.4: Fragmento de código en el que se calculan los porcentajes de reducción de etiquetas y del total de caracteres.

```
System.out.println("\n"+"Archivo Minimizado "+z+": "
      + rutaficherominimizado);
System.out.println("El archivo tiene "
      + caracteresEtiquetasMinimizado
      +" caracteres de etiquetas y "
      + caracteresMinimizado +" caracteres en total.");

System.out.println("Los caracteres de etiquetas "
      +"se han reducido un "+ df.format(reduccionEtiquetas)
      + " % y los caracteres totales un "
      + df.format(reduccionArchivo)+" %.");
```

Figura 4.5: Fragmento de código en el que se muestran los resultados por pantalla.

En el Anexo B.2 se incluye el código completo correspondiente a esta clase.

4.1.3. Salida del programa

A la hora de ejecutar el programa se le ha pasado como argumento el directorio C:\sgml\pruebas_bbdd. En dicho directorio se han incluido 24 archivos, de los cuales 3 son archivos sin minimizar que contienen 1, 5 y 50 registros de la base de datos, y el resto, son archivos obtenidos al aplicar distintas minimizaciones a los primeros. Se incluye, a continuación, una lista con los nombres de los archivos utilizados y el documento del que proceden.

- **registro1.sgml**: Archivo original que contiene un registro de la base de datos.

  Archivos procedentes de registro1.sgml:

  - **registro1_min_omit1.sgml**
  - **registro1_min_omit2.sgml**
  - **registro1_min_omit3.sgml**
  - **registro1_min_short1.sgml**
  - **registro1_min_short2.sgml**
  - **registro1_min_short3.sgml**
  - **registro1_min_short4.sgml**
  - **registro1_min_short5.sgml**
  - **registro1_min_short6.sgml**
  - **registro1_min_todo.sgml**

- **registro5.sgml**: Archivo original que contiene cinco registros de la base de datos.

  Archivos procedentes de registro5.sgml:

  - **registro5_min_omit1.sgml**
  - **registro5_min_omit2.sgml**
  - **registro5_min_omit3.sgml**
  - **registro5_min_short1.sgml**
  - **registro5_min_short2.sgml**
  - **registro5_min_short3.sgml**
  - **registro5_min_short4.sgml**
  - **registro5_min_short5.sgml**
  - **registro5_min_short6.sgml**
  - **registro5_min_todo.sgml**

- **registro50.sgml**: Archivo original que contiene cincuenta registros de la base de datos.

  Archivo procedente de registro50.sgml:

  - **registro50_min_todo.sgml**

Una vez ejecutado el programa se obtiene la salida que se muestra en Figura 4.6. Los resultados serán analizados en la Sección 4.2.

Archivo completo: C:\sgml\pruebas_bbdd\registro1.sgml
El archivo tiene 313 caracteres de etiquetas y 370 caracteres en total.

Archivos obtenidos al minimizar registro1.sgml:

Archivo Minimizado 1: C:\sgml\pruebas_bbdd\registro1_min_omit1.sgml
El archivo tiene 185 caracteres de etiquetas y 242 caracteres en total.
Los caracteres de etiquetas se han reducido un 40,89 % y los caracteres totales un 34,59 %.

Archivo Minimizado 2: C:\sgml\pruebas_bbdd\registro1_min_omit2.sgml
El archivo tiene 244 caracteres de etiquetas y 301 caracteres en total.
Los caracteres de etiquetas se han reducido un 22,04 % y los caracteres totales un 18,65 %.

Archivo Minimizado 3: C:\sgml\pruebas_bbdd\registro1_min_omit3.sgml
El archivo tiene 137 caracteres de etiquetas y 194 caracteres en total.
Los caracteres de etiquetas se han reducido un 56,23 % y los caracteres totales un 47,57 %.
Archivo Minimizado 4: C:\sgml\pruebas_bbdd\registro1_min_short1.sgml
El archivo tiene 286 caracteres de etiquetas y 343 caracteres en total.
Los caracteres de etiquetas se han reducido un 8,63 % y los caracteres totales un 7,30 %.

Archivo Minimizado 5: C:\sgml\pruebas_bbdd\registro1_min_short2.sgml
El archivo tiene 239 caracteres de etiquetas y 296 caracteres en total.
Los caracteres de etiquetas se han reducido un 23,64 % y los caracteres totales un 20,00 %.

Archivo Minimizado 6: C:\sgml\pruebas_bbdd\registro1_min_short3.sgml
El archivo tiene 203 caracteres de etiquetas y 260 caracteres en total.
Los caracteres de etiquetas se han reducido un 35,14 % y los caracteres totales un 29,73 %.

Archivo Minimizado 7: C:\sgml\pruebas_bbdd\registro1_min_short4.sgml
El archivo tiene 305 caracteres de etiquetas y 362 caracteres en total.
Los caracteres de etiquetas se han reducido un 2,56 % y los caracteres totales un 2,16 %.

Archivo Minimizado 8: C:\sgml\pruebas_bbdd\registro1_min_short5.sgml
El archivo tiene 291 caracteres de etiquetas y 348 caracteres en total.
Los caracteres de etiquetas se han reducido un 7,03 % y los caracteres totales un 5,95 %.

Archivo Minimizado 9: C:\sgml\pruebas_bbdd\registro1_min_short6.sgml
El archivo tiene 179 caracteres de etiquetas y 236 caracteres en total.
Los caracteres de etiquetas se han reducido un 42,81 % y los caracteres totales un 36,22 %.

Archivo Minimizado 10: C:\sgml\pruebas_bbdd\registro1_min_todo.sgml
El archivo tiene 106 caracteres de etiquetas y 163 caracteres en total.
Los caracteres de etiquetas se han reducido un 66,13 % y los caracteres totales un 55,95 %.

Archivo completo: C:\sgml\pruebas_bbdd\registro5.sgml
El archivo tiene 1414 caracteres de etiquetas y 1730 caracteres en total.

Archivos obtenidos al minimizar registro5.sgml:

Archivo Minimizado 1: C:\sgml\pruebas_bbdd\registro5_min_omit1.sgml
El archivo tiene 763 caracteres de etiquetas y 1079 caracteres en total.
Los caracteres de etiquetas se han reducido un 46,04 % y los caracteres totales un 37,63 %.

Archivo Minimizado 2: C:\sgml\pruebas_bbdd\registro5_min_omit2.sgml
El archivo tiene 1161 caracteres de etiquetas y 1477 caracteres en total.
Los caracteres de etiquetas se han reducido un 17,89 % y los caracteres totales un 14,62 %.

Archivo Minimizado 3: C:\sgml\pruebas_bbdd\registro5_min_omit3.sgml
El archivo tiene 555 caracteres de etiquetas y 871 caracteres en total.
Los caracteres de etiquetas se han reducido un 60,75 % y los caracteres totales un 49,65 %.

Archivo Minimizado 4: C:\sgml\pruebas_bbdd\registro5_min_short1.sgml
El archivo tiene 1284 caracteres de etiquetas y 1600 caracteres en total.
Los caracteres de etiquetas se han reducido un 9,19 % y los caracteres totales un 7,51 %.

Archivo Minimizado 5: C:\sgml\pruebas_bbdd\registro5_min_short2.sgml
El archivo tiene 1024 caracteres de etiquetas y 1340 caracteres en total.
Los caracteres de etiquetas se han reducido un 27,58 % y los caracteres totales un 22,54 %.

Archivo Minimizado 6: C:\sgml\pruebas_bbdd\registro5_min_short3.sgml
El archivo tiene 850 caracteres de etiquetas y 1166 caracteres en total.
Los caracteres de etiquetas se han reducido un 39,89 % y los caracteres totales un 32,60 %.

Archivo Minimizado 7: C:\sgml\pruebas_bbdd\registro5_min_short4.sgml
El archivo tiene 1376 caracteres de etiquetas y 1692 caracteres en total.
Los caracteres de etiquetas se han reducido un 2,69 % y los caracteres totales un 2,20 %.

Archivo Minimizado 8: C:\sgml\pruebas_bbdd\registro5_min_short5.sgml
El archivo tiene 1308 caracteres de etiquetas y 1624 caracteres en total.
Los caracteres de etiquetas se han reducido un 7,50 % y los caracteres totales un 6,13 %.

Archivo Minimizado 9: C:\sgml\pruebas_bbdd\registro5_min_short6.sgml
El archivo tiene 734 caracteres de etiquetas y 1050 caracteres en total.
Los caracteres de etiquetas se han reducido un 48,09 % y los caracteres totales un 39,31 %.

Archivo Minimizado 10: C:\sgml\pruebas_bbdd\registro5_min_todo.sgml
El archivo tiene 418 caracteres de etiquetas y 734 caracteres en total.
Los caracteres de etiquetas se han reducido un 70,44 % y los caracteres totales un 57,57 %.

Archivo completo: C:\sgml\pruebas_bbdd\registro50.sgml
El archivo tiene 28916 caracteres de etiquetas y 35608 caracteres en total.

Archivos obtenidos al minimizar registro50.sgml:

Archivo Minimizado 1: C:\sgml\pruebas_bbdd\registro50_min_todo.sgml
El archivo tiene 9205 caracteres de etiquetas y 15897 caracteres en total.
Los caracteres de etiquetas se han reducido un 68,17 % y los caracteres totales un 55,36 %.

Figura 4.6: Salida del programa.

4.2. Comparación de los resultados

Las mediciones realizadas pretenden determinar el tanto por ciento de reducción de etiquetas conseguida en cada archivo estudiado. Con ese fin, en el apartado anterior, se han contado los caracteres totales y los caracteres correspondientes a etiquetas de cada archivo. En este apartado se resumen los resultados obtenidos en forma de tablas y se comparan mediante gráficas.

Las comparaciones se hacen, en primer lugar, entre las distintas opciones de cada característica de minimización, para determinar si el uso aislado de algunas de ellas puede ser ventajoso según el tipo de objetivo que se desee conseguir, por ejemplo, si se busca un equilibrio entre reducción de tamaño y facilidad de aplicación. Posteriormente,

se muestran los resultados obtenidos para la combinación de minimizaciones que consigue mayor reducción de tamaño, y se compara con los mejores resultados obtenidos usando solo Omittag o Shorttag. Para finalizar, se comprueba la dependencia de los resultados con el número de registros considerados.

4.2.1. Documentos sin minimizar

Los documentos sin minimizar con los que se trabaja son, por un lado, el documento del Anexo A.1 (registro50.sgml), que contiene la información correspondiente a cincuenta registros de la base de datos, y por otro lado, dos extractos de ese mismo documento de uno y cinco registros (registro1.sgml y registro5.sgml). El archivo registro5.sgml fue usado en el capítulo anterior para probar distintos tipos de minimización y escoger la combinación óptima que se aplica en el documento de más longitud. En este capítulo se sigue haciendo uso de los archivos obtenidos a partir del mismo, para comparar las reducciones conseguidas según el tipo de minimización aplicada. Además, se usan archivos minimizados con un solo registro para estudiar la dependencia con el número de registros considerados.

En la Tabla 4.1 se han incluido los resultados del programa referidos a caracteres de etiquetas y caracteres totales de los archivos sin minimizar. Como se esperaba, el porcentaje que suponen las etiquetas respecto al total del archivo es muy elevado. En concreto, las etiquetas suponen un 85 %, un 82 % y un 81 % del total de su tamaño para los archivos registro1.sgml, registro5.sgml y registro50.sgml respectivamente. Las diferencias no sólo se deben a la longitud de los datos almacenados en los registros, también varía el número de elementos de cada uno, pues hay muchos elementos opcionales. Aún así son porcentajes muy similares.

Archivos sin minimizar	Caracteres de etiquetas	Caracteres totales
registro1.sgml	313	370
registro5.sgml	1414	1730
registro50.sgml	28916	35608

Tabla 4.1: Caracteres de etiquetas y caracteres totales de los archivos completos con los que se trabaja.

4.2.2. Resultados obtenidos con Omittag

La aplicación de Omittag ha permitido eliminar muchas de las etiquetas, de inicio y de fin, en los documentos con los que se ha trabajado. Se ha decidido llamar Omittag1 (así aparecerá en las gráficas) a la posibilidad de omitir etiquetas de fin. Esta opción se ha usado en los archivos registro1_min_omit1.sgml y registro5_min_omit1.sgml (el segundo puede verse en la Figura 3.40 y el primero es un extracto del mismo con un solo registro). De igual manera, se llamará Omittag2 a la opción de omitir

las etiquetas de inicio cuando sea posible, como en registro1_min_omit2.sgml y registro5_min_omit2.sgml y Omittag3 a la omisión de todas las etiquetas de inicio y fin posibles en un documento, que es el caso de registro1_min_omit3.sgml y registro5_min_omit3.sgml.

En la Tablas 4.2 y 4.3 se resumen los resultados de las mediciones hechas a los archivos mencionados.

Archivos con 1 registro	Caracteres de etiquetas	Caracteres totales	Reducción de caracteres de etiquetas (%)	Reducción de caracteres totales (%)
registro1_min_omit1.sgml	185	242	40,89	34,59
registro1_min_omit2.sgml	244	301	22,04	18,65
registro1_min_omit3.sgml	137	194	56,23	47,57

Tabla 4.2: Resultados obtenidos para un registro minimizado con Omittag.

Archivos con 5 registros	Caracteres de etiquetas	Caracteres totales	Reducción de caracteres de etiquetas (%)	Reducción de caracteres totales (%)
registro5_min_omit1.sgml	763	1079	46,04	37,63
registro5_min_omit2.sgml	1161	1477	17,89	14,62
registro5_min_omit3.sgml	555	871	60,75	49,65

Tabla 4.3: Resultados obtenidos para cinco registros minimizados con Omittag.

A partir de estos resultados se decide generar varias gráficas que ilustren mejor la comparación que se quiere realizar. En las dos primeras gráficas (Figuras 4.7 y 4.8) se muestra el número de caracteres de etiquetas antes y después de las minimizaciones consideradas. La primera corresponde a los resultados obtenidos al considerar un registro y en la segunda se consideran cinco registros de la base de datos.

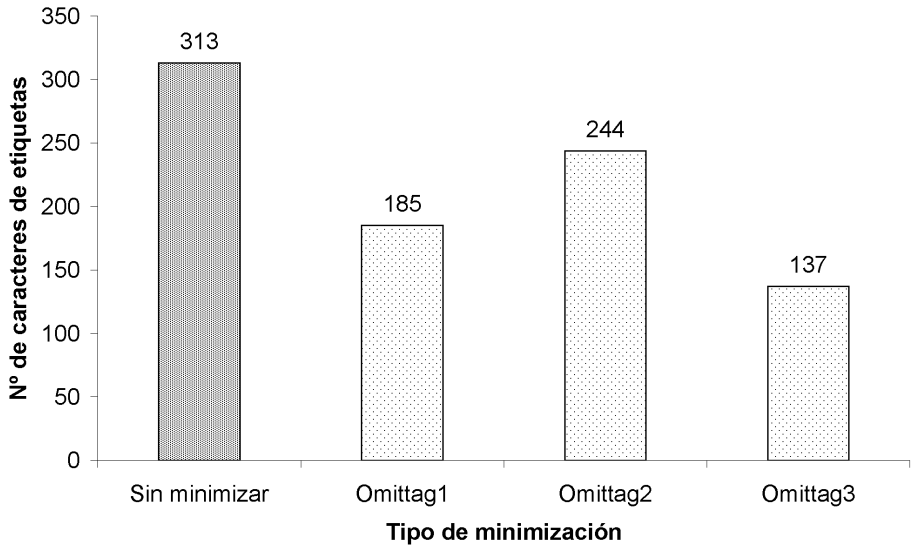

Figura 4.7: Número de caracteres de etiquetas de un registro minimizado con Omittag.

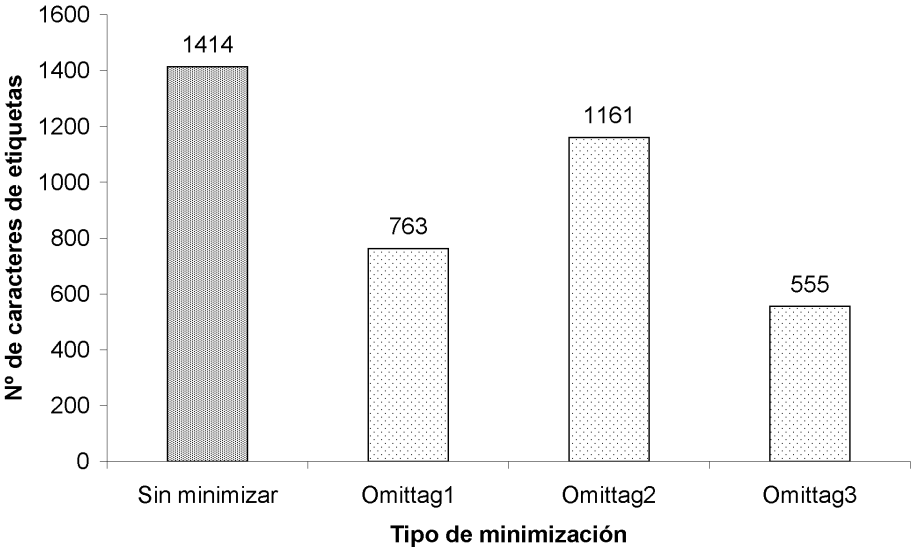

Figura 4.8: Número de caracteres de etiquetas de cinco registros minimizados con Omittag.

Como era de esperar, el archivo en el que se combinan ambas posibilidades, la omisión de etiquetas de inicio y de fin, es el que más se ha reducido, sin embargo, es interesante observar la gran reducción conseguida y las ventajas que supone el uso de la primera opción de forma independiente. Con dicha opción pueden eliminarse todas las etiquetas de fin, lo que supone una reducción del 50 % de las etiquetas del archivo, y su aplicación es muy sencilla (pudo verse al realizar la minimización en la sección 3.3.3.1). Las etiquetas de inicio eliminadas son bastantes menos y, decidir cuáles pueden ser omitidas, aunque no es complicado, lleva algo más de tiempo. Las etiquetas de inicio que pueden minimizarse al usar ambas opciones a la vez son aún menos, sin embargo, la diferencia entre los

caracteres de etiquetas de las opciones Omittag3 y Omittag1 sigue siendo notable como para dejar de considerar esta opción.

Se incluyen dos gráficas con los resultados en forma de porcentajes (Figuras 4.9 y 4.10). Con ellas se pretende facilitar la comparación entre archivos de distinto tamaño.

Se debe recordar que a la hora de contar caracteres de etiquetas se decidió sumar también los caracteres de la declaración DOCTYPE, pues aunque no pueden minimizarse, son parte del marcado, por eso los resultados para Omit1 son inferiores al 50 % que se esperaba.

Los resultados obtenidos para la combinación de minimizaciones son de un 56 % de reducción de caracteres de etiquetas, si se considera un registro, y de más de un 60 % para cinco registros. La diferencia es debida, en cierta medida, al peso de la declaración DOCTYPE, que es menor en el segundo caso, y a que en el mismo han podido omitirse más etiquetas de inicio que en el primero.

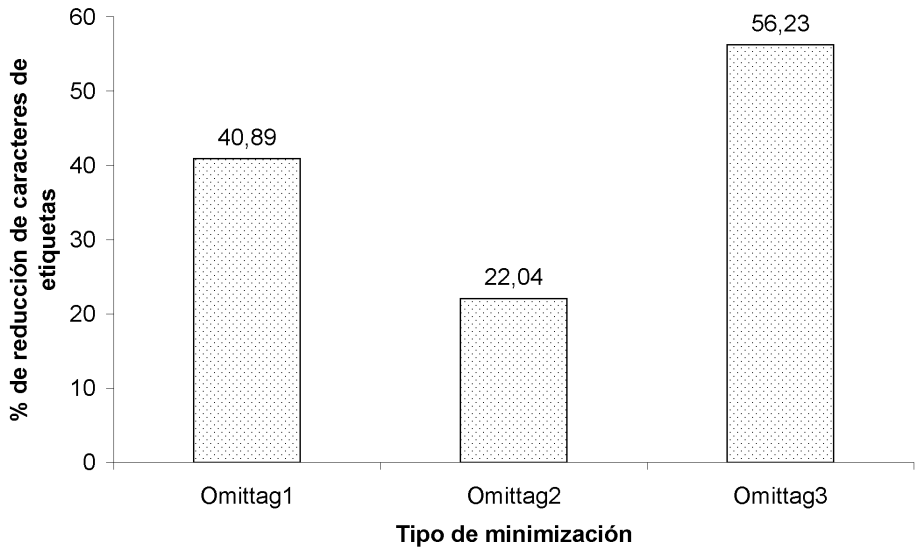

Figura 4.9: Porcentaje de reducción de caracteres de etiquetas para un registro minimizado con Omittag.

Ese 60 % de reducción de caracteres de etiquetas supone que el archivo ha quedado reducido casi a la mitad de su longitud inicial, lo que lleva a pensar de nuevo, en la cantidad tan grande de etiquetas que necesita un archivo de este tipo para almacenar pocos datos. Por lo tanto, aunque no pueden generalizarse los resultados, pues dependen de la densidad de etiquetas y de la estructura del documento, puede verse que, en archivos de este tipo, la reducción de tamaño es considerable. Además, si se considera eliminar sólo las etiquetas de fin de elemento, no existe una gran dificultad a la hora de minimizar ni se complica la legibilidad del documento (más bien al contrario, como puede verse en la Figura 3.40).

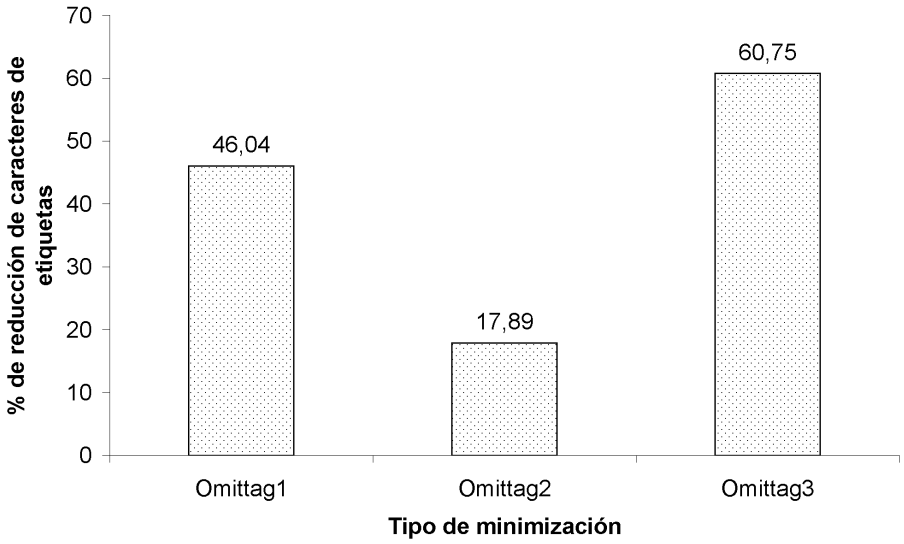

Figura 4.10: Porcentaje de reducción de caracteres de etiquetas para cinco registros minimizados con Omittag.

4.2.3. Resultados obtenidos con Shorttag

Las distintas alternativas de Shorttag, a la hora de reducir el tamaño de una etiqueta, se usaron en el capítulo anterior para obtener seis archivos minimizados diferentes, partiendo de un mismo archivo completo. Dichos archivos se corresponden con las siguientes reducciones:

- Shorttag1: Se elimina el carácter tagc en los archivos registro1_min_short1.sgml y registro5_min_short1.sgml.
- Shorttag2: Se elimina el identificador de las etiquetas de fin en los archivos registro1_min_short2.sgml y registro5_min_short2.sgml.
- Shorttag3: Se usa NET en los archivos registro1_min_short3.sgml y registro5_min_short3.sgml.
- Shorttag4: Se eliminan las comillas en los valores de atributos de los archivos registro1_min_short4.sgml y registro5_min_short4.sgml.
- Shorttag5: Se usa sólo el valor del atributo (siempre que esté incluído en la DTD) en los archivos registro1_min_short5.sgml y registro5_min_short5.sgml.
- Shorttag6: Se usan conjuntamente todas las posibilidades anteriores en los archivos registro1_min_short6.sgml y registro5_min_short6.sgml.

Las Tablas 4.4 y 4.5 contienen los resultados de las mediciones hechas a dichos archivos.

Archivos con 1 registro	Caracteres de etiquetas	Caracteres totales	Reducción de caracteres de etiquetas (%)	Reducción de caracteres totales (%)
registro1_min_short1.sgml	286	343	8,63	7,3
registro1_min_short2.sgml	239	296	23,64	20
registro1_min_short3.sgml	203	260	35,14	29,73
registro1_min_short4.sgml	305	362	2,56	2,16
registro1_min_short5.sgml	291	348	7,03	5,95
registro1_min_short6.sgml	179	236	42,81	36,22

Tabla 4.4: Resultados obtenidos para un registro minimizado con Shorttag.

Archivos con 5 registros	Caracteres de etiquetas	Caracteres totales	Reducción de caracteres de etiquetas (%)	Reducción de caracteres totales (%)
registro5_min_short1.sgml	1284	1600	9,19	7,51
registro5_min_short2.sgml	1024	1340	27,58	22,54
registro5_min_short3.sgml	850	1166	39,89	32,6
registro5_min_short4.sgml	1376	1692	2,69	2,2
registro5_min_short5.sgml	1308	1624	7,5	6,13
registro5_min_short6.sgml	734	1050	48,09	39,31

Tabla 4.5: Resultados obtenidos para cinco registros minimizados con Shorttag.

Con estos resultados se generan dos gráficas (Figuras 4.11 y 4.12) en las que se compara el número de caracteres de etiquetas que tienen los archivos antes y después de ser minimizados.

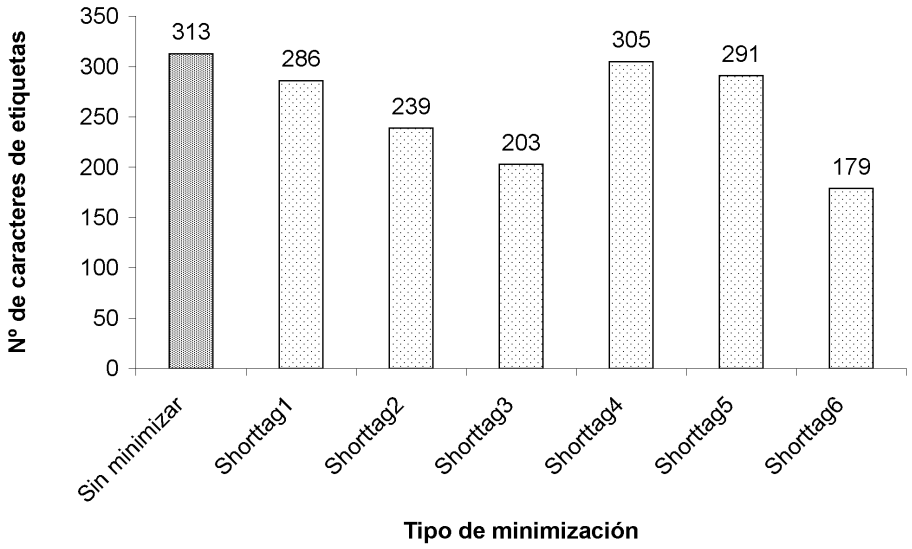

Figura 4.11: Número de caracteres de etiquetas de un registro minimizado con Shorttag.

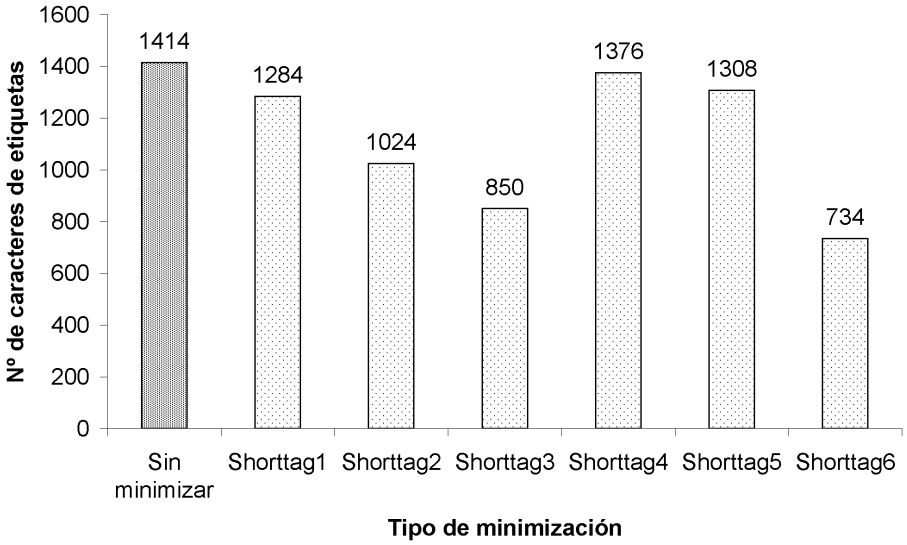

Figura 4.12: Número de caracteres de etiquetas de cinco registros minimizados con Shorttag.

Las reducciones de etiquetas obtenidas son menores que en el caso de Omittag, pues en ningún caso se eliminan las etiquetas por completo, sino partes de las mismas, sin embargo, no son reducciones pequeñas, y deben tenerse en cuenta las posibles ventajas de Shorttag frente a Omittag [47]. Una ventaja de Shorttag es que para usar muchas de sus opciones, por ejemplo la omisión del identificador en las etiquetas de fin, no es necesario incluir una DTD, mientras que el uso de Omittag sin DTD puede dar lugar a ambigüedades. Por otro lado, la omisión de etiquetas puede facilitar que se cometan más errores en la estructura de los documentos (como los vistos en la Sección 3.2.1), aunque no parece el caso de las bases de datos.

Para hacerse una idea más clara de las reducciones obtenidas se incluyen en forma de porcentajes en las Figuras 4.13 y 4.14.

Es interesante fijarse en que sólo con la eliminación del carácter tagc de algunas etiquetas, que fue una de las propuestas para MicroXML[7], se ha conseguido una reducción cercana al 9 %. En cuanto a las minimizaciones relativas a los atributos, se debe señalar que no proporcionan grandes reducciones debido a la escasez de atributos del ejemplo utilizado.

Los mejores resultados se consiguen con la eliminación de los identificadores y el uso de NET, casi un 28 % y un 40 % de reducción de los caracteres de etiquetas (en el archivo con cinco registros) respectivamente, que no son cantidades despreciables y aún menos usadas conjuntamente con las anteriores, donde se llega al 48 % de las etiquetas, que supone más de un 39 % del archivo. En una base de datos de 1GB estaríamos hablando de más de 390 MB reducidos.

Lo cierto es que el uso de NET complica bastante la lectura del documento pero eliminar el identificador en la etiquetas de fin en este tipo de documentos, como ha podido verse en el apartado 3.3.4.2, no tiene gran dificultad para el usuario, ni complica la lectura, ni parece que complique significativamente el parser [46]. Limitándose a este tipo de minimización se ha conseguido reducir el archivo en un 22 % de su tamaño original (20 % en el caso de un solo registro).

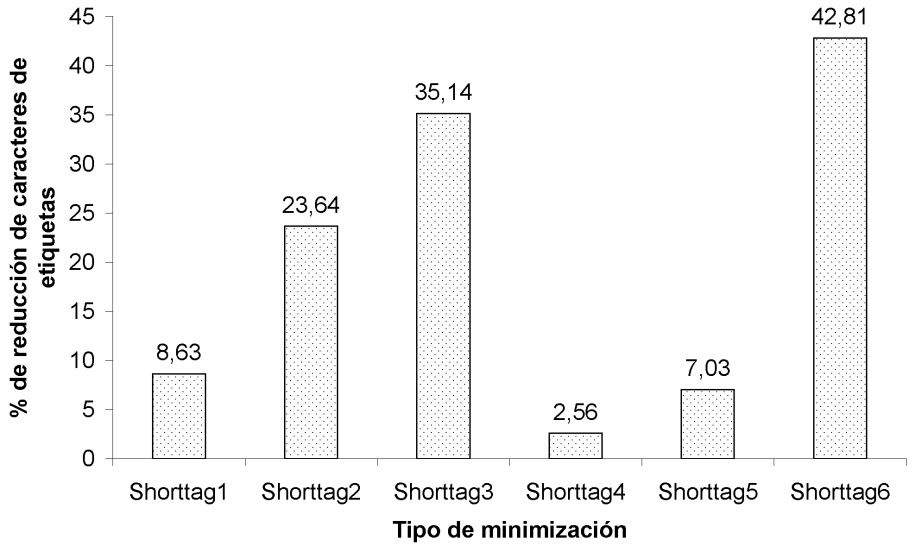

Figura 4.13: Porcentaje de reducción de caracteres de etiquetas para un registro minimizado con Shorttag.

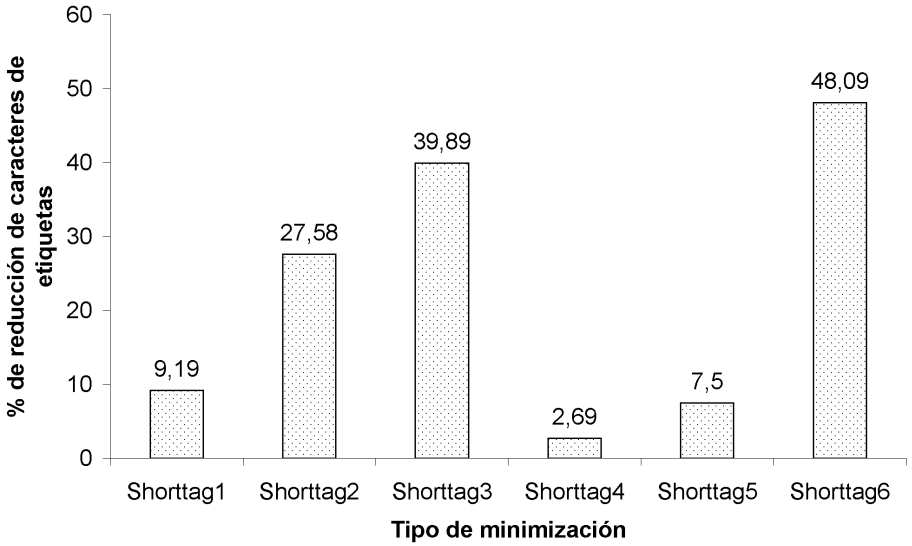

Figura 4.14: Porcentaje de reducción de caracteres de etiquetas para cinco registros minimizados con Shorttag.

4.2.4. Resultados obtenidos con Omittag y Shorttag conjuntamente

La combinación de minimizaciones, con la que más se ha conseguido reducir los archivos iniciales, dio como resultado los archivos registro1_min_todo.sgml, registro5_min_todo.sgml y registro50_min_todo.sgml, al considerar 1, 5 y 50 registros respectivamente. Los resultados de dicha minimización se resumen en la Tabla 4.6.

Para ilustrar la reducción de caracteres de etiquetas conseguida se generan las gráficas que aparecen en las Figuras 4.15, 4.16, 4.17.

La aplicación de minimizaciones es más laboriosa en este caso, sin embargo, basta comparar la extensión de los archivos mostrados en las Figuras A.1 y A.2 para darse cuenta de la gran reducción conseguida. Se han eliminado aproximadamente un 68 % de los caracteres de etiquetas con lo que se consigue una reducción de archivo en torno al 55 %. Es interesante apreciar que la diferencia entre la reducción conseguida aquí y la conseguida usando sólo Omittag (para el caso de cinco registros) es de casi un 10 % adicional, teniendo en cuenta que al usar Omittag no se usan las características de Shorttag que más reducción aportaban y, que es un documento con muy pocos atributos, es un buen resultado. La diferencia no es despreciable y lo sería aún menos si los atributos tuviesen más peso en el documento.

Archivos	Caracteres de etiquetas	Caracteres totales	Reducción de caracteres de etiquetas (%)	Reducción de caracteres totales (%)
registro1_min_todo.sgml	106	163	66,13	55,95
registro5_min_todo.sgml	418	734	70,44	57,57
registro50_min_todo.sgml	9205	15897	68,17	55,36

Tabla 4.6: Resultados obtenidos para uno, cinco y cincuenta registros minimizados con Omittag y Shorttag.

Figura 4.15: Número de caracteres de etiquetas de un registro minimizado con Omittag y Shorttag.

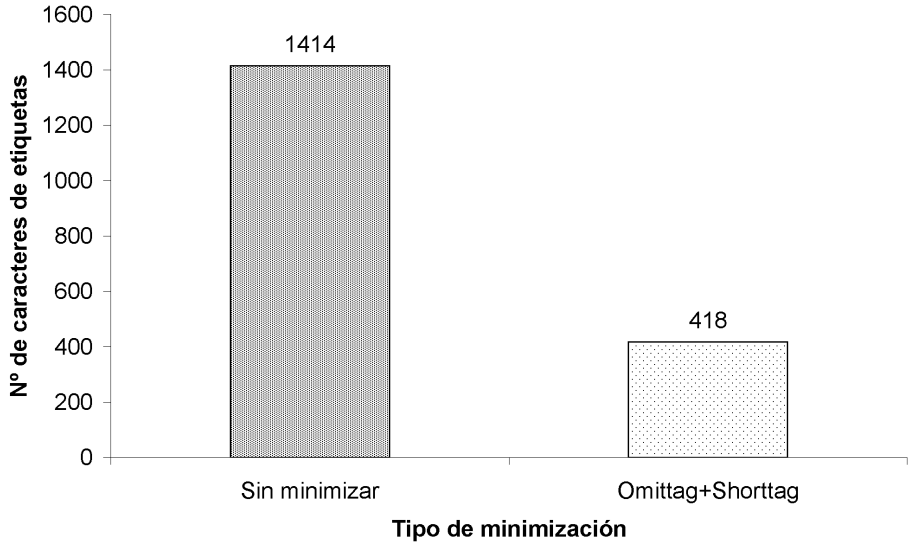

Figura 4.16: Número de caracteres de etiquetas de cinco registros minimizados con Omittag y Shorttag.

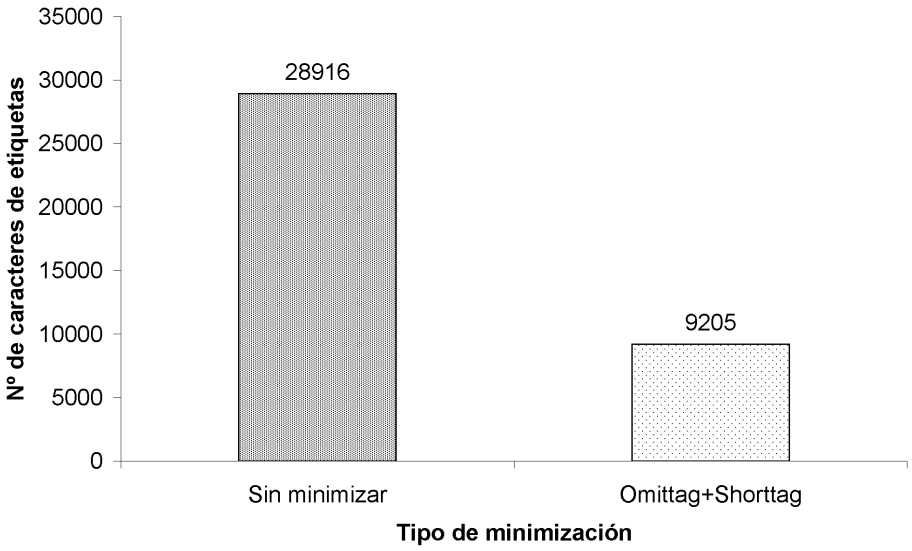

Figura 4.17: Número de caracteres de etiquetas de cincuenta registros minimizados con Omittag y Shorttag.

Ya se ha comentado que, en documentos de este tipo, las etiquetas suponen un porcentaje importante del tamaño de archivo. Puede ser ilustrativo ver gráficamente cómo varía el porcentaje que suponen las etiquetas antes y después de minimizar, por ello, se han incluído las gráficas de las Figuras 4.18, 4.19 y 4.20.

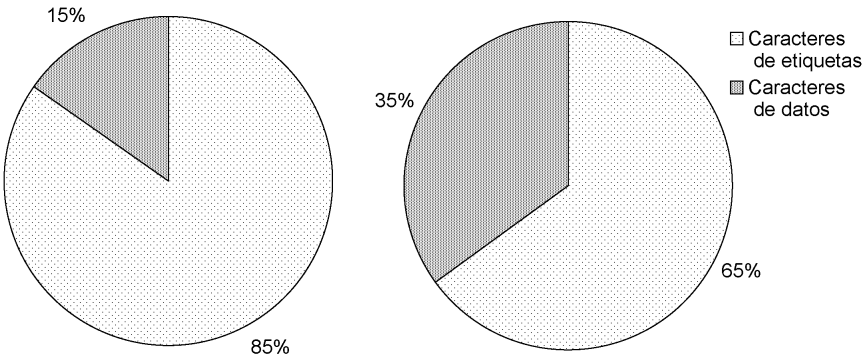

Figura 4.18: Porcentaje de etiquetas y datos para un registro antes (1ª figura) y después (2ª figura) de minimizarlo con Omittag y Shorttag.

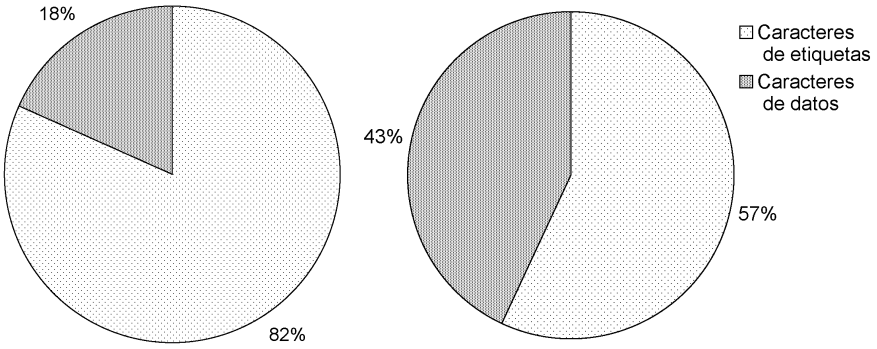

Figura 4.19: Porcentaje de etiquetas y datos para cinco registros antes (1ª figura) y después (2ª figura) de minimizarlos con Omittag y Shorttag.

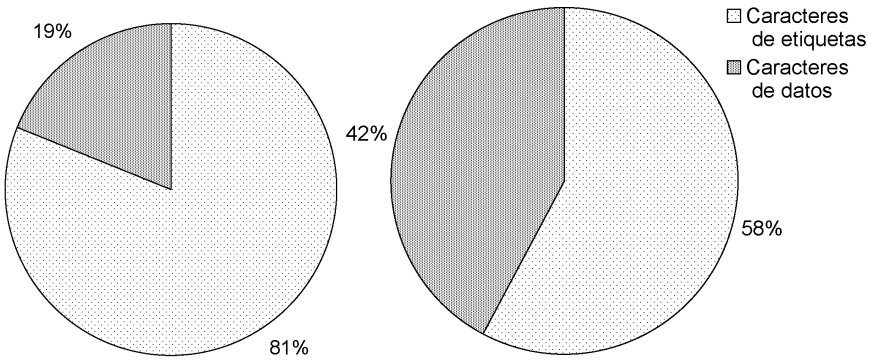

Figura 4.20: Porcentaje de etiquetas y datos para cincuenta registros antes (1ª figura) y después (2ª figura) de minimizarlos con Omittag y Shorttag.

En estas gráficas se observa, además, que a pesar de la estructura tan variable que tienen los registros de la base de datos utilizada (por tener muchos elementos opcionales), los porcentajes son similares.

4.2.5. Comparativa según las minimizaciones empleadas y el número de registros considerados

Aunque ya se han comparado entre sí los resultados obtenidos para las distintas posibilidades de Omittag y las de Shorttag, se incluye aquí una nueva gráfica (Figura 4.21) para remarcar las reducciones totales de archivo obtenidas con la mejor combinación de opciones de Omittag, las de Shorttag y la aplicación de todas las posibilidades.

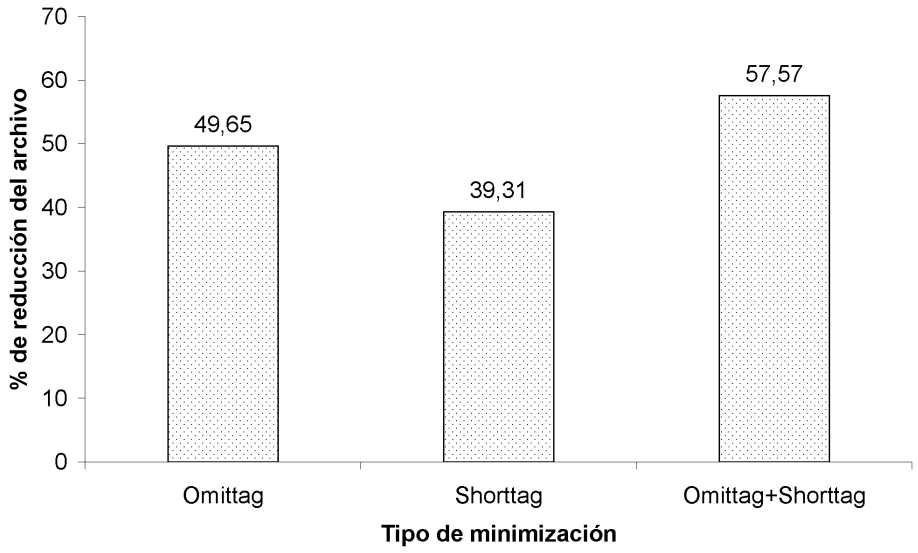

Figura 4.21: Porcentaje de reducción de caracteres para cinco registros minimizados con todas las posibilidades de Omittag, las de Shorttag y las de ambas a la vez.

Puede verse, nuevamente, que los resultados obtenidos con Omittag son mejores que con Shorttag y, que al usar ambas opciones, se consigue una mejora no despreciable de los resultados, a pesar de que no se usan las posibilidades de minimización más ventajosas de Shorttag (al haber omitido todas las etiquetas de fin, no pueden usarse ni net, ni la eliminación del identificador de dichas etiquetas).

Por último, se compara la dependencia de los resultados con el número de registros considerados en las Figuras 4.22 y 4.23. Puede apreciarse que no hay grandes diferencias, aunque los resultados obtenidos con cinco registros son algo mejores. Se debe, en cierta medida, a que el porcentaje de etiquetas de dicho archivo es algo mayor que en el de cincuenta registros. En el caso de un registro, los resultados son peores por haber considerado la declaración DOCTYPE como parte de las etiquetas (se ha considerado así en todos los casos, pero en el de un registro esa declaración supone un tanto por ciento más alto del total de archivo.)

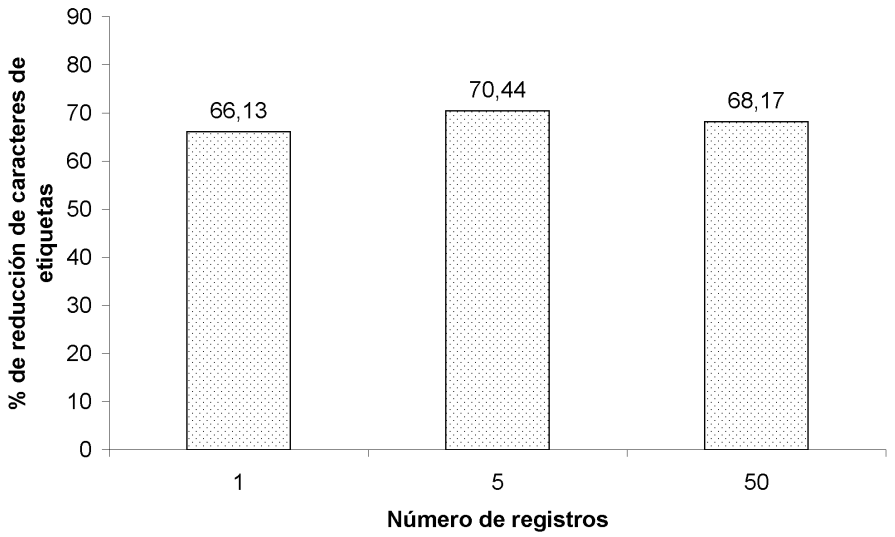

Figura 4.22: Porcentaje de reducción de etiquetas conseguido para uno, cinco y cincuenta registros.

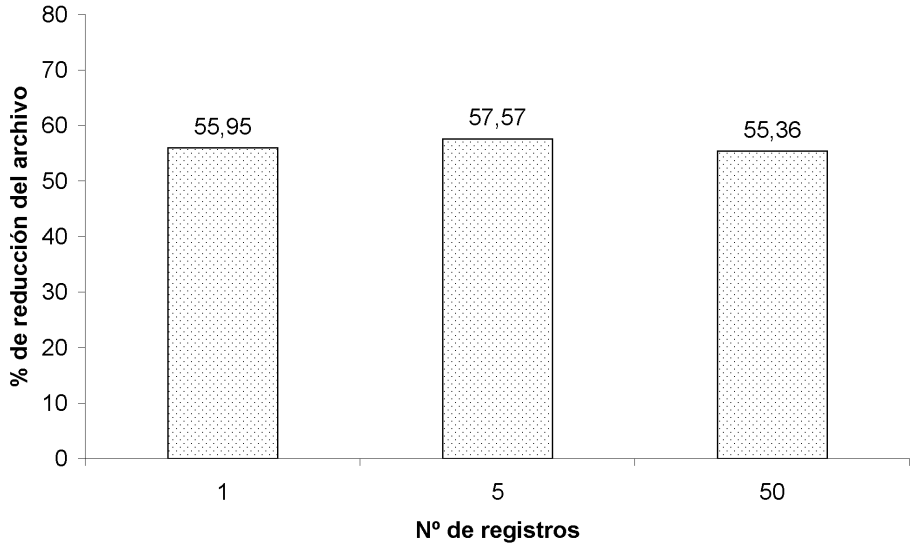

Figura 4.23: Porcentaje de reducción de caracteres totales conseguidos para uno, cinco y cincuenta registros.

4.3. Conclusiones

En este capítulo se ha descrito el programa desarrollado para obtener el porcentaje de reducción de caracteres de etiquetas conseguido al minimizar los documentos con los que se trabaja. Posteriormente, se han expuesto dichos resultados en forma de gráficas. Las gráficas muestran, además, el porcentaje que representan dichas etiquetas en el total de archivo, antes y después de las minimizaciones. En ellas se aprecia que se ha podido conseguir una reducción de los caracteres correspondientes a etiquetas del archivo

completo de un 68 %, lo que ha supuesto una reducción total del documento de más del 55 %. En el siguiente capítulo se verán las conclusiones globales de este proyecto.

Capítulo 5

Conclusiones y líneas futuras

El objetivo principal de este proyecto ha sido estudiar las características de minimización del lenguaje SGML, concretamente, medir la reducción que experimenta un documento SGML al hacer uso de dichas características y comprobar la dificultad que supone su aplicación para un usuario del lenguaje. El interés por este tema surge con el nacimiento de un nuevo lenguaje de marcado, MicroXML, una nueva simplificación de XML, que a su vez aparece como simplificación de SGML. Con el desarrollo de los primeros borradores de MicroXML, comienzan los debates sobre qué aspectos de XML deben mantenerse, pero también otros sobre la posibilidad de rescatar algunas de las opciones de SGML eliminadas en XML, entre ellas ciertas características de minimización. Se quiere determinar en qué medida podría ser beneficioso para algunos usuarios el hecho de mantener este tipo de aspectos opcionales en futuros lenguajes de marcado.

El proyecto incluye un estudio previo del concepto de lenguaje de marcado y los lenguajes SGML, XML y MicroXML, necesario para aprender a trabajar con archivos SGML y comprobar la forma en que se ha ido simplificando el lenguaje. En dicho estudio se ha revisado la norma ISO 8879:1986 (SGML), la recomendación del W3C para XML y la última especificación propuesta para MicroXML (la última en el momento en que se escribe este proyecto), apreciándose con ello la diferente dificultad existente a la hora de acercarse a estos lenguajes. SGML se consideró en su momento demasiado complejo para los usuarios y para el tipo de información que suele usarse en la Web y ha ocurrido lo mismo con XML. A pesar de la cantidad de herramientas ya existentes para trabajar con XML, se decide volver a simplificarlo aceptando la pérdida de flexibilidad que ello supone.

Para conocer las diferencias entre estos lenguajes, en este proyecto se han detallado las partes que forman los documentos SGML y XML, así como la sintaxis de sus respectivas DTDs. En cuanto a MicroXML, también se ha probado uno de los primeros analizadores desarrollados. El ejemplo utilizado para probar el analizador se escribió como documento SGML y se fue transformando para adaptarlo a XML y finalmente a MicroXML.

Por otro lado, se ha hecho un estudio de las características de minimización descritas para SGML. Como parte del mismo, se han aplicado las minimizaciones a un ejemplo sencillo, lo que ha servido para determinar la forma correcta de utilizarlas y los problemas que pueden encontrarse con su uso. Además, para trabajar con los archivos creados y

comprobar la validez de las minimizaciones aplicadas, se han buscado los analizadores disponibles para SGML y se ha instalado uno de ellos (OpenSP) junto con un editor.

Una vez conocida la forma de aplicar las minimizaciones, se ha escogido un ejemplo entre los tipos de documentos que más pueden beneficiarse del uso de dichas opciones de SGML, aquellos que tienen un porcentaje elevado de etiquetas respecto al total del tamaño de archivo. En este caso, el documento elegido contiene cincuenta registros de una base de datos. Se ha usado un extracto más manejable de este archivo para aplicar, una a una, las opciones de minimización que han resultado adecuadas para el ejemplo. De esta forma se ha podido valorar la dificultad encontrada al aplicarlas y si complican, o no, la lectura de los documentos. Después se ha determinado la combinación óptima de dichas minimizaciones, atendiendo sólo a la reducción de tamaño conseguida, y se ha aplicado al archivo completo. Todos los archivos minimizados han sido analizados con OpenSP para comprobar la validez de las minimizaciones realizadas. Con todo ello, se han generado diez archivos minimizados distintos partiendo de un mismo archivo que contiene un registro de la base de datos, otros diez partiendo de un archivo que contiene cinco registros y, finalmente, un archivo minimizado para el que contiene cincuenta registros.

A la hora de obtener resultados numéricos para estos archivos, se decide desarrollar un programa en lenguaje Java que facilite la tarea. El programa lee los archivos contenidos en un directorio que se le pasa como argumento, asocia cada archivo completo con sus archivos minimizados y devuelve los resultados por pantalla. Concretamente, muestra el número de caracteres totales y de etiquetas de todos ellos, además del tanto por ciento de reducción de caracteres totales y de etiquetas en los minimizados.

Los resultados obtenidos con el programa se han expuesto en forma de gráficas para facilitar su análisis. En ellas se aprecia que se ha podido conseguir una reducción de los caracteres correspondientes a etiquetas del archivo completo de un 68 %, lo que ha supuesto una reducción total del documento de más del 55 %. Es decir, el documento ha quedado reducido a menos de la mitad de su extensión original. Visto de otra forma, si los datos del documento antes de la minimización suponían sólo el 20 % del total (el resto era marcado), tras la minimización los datos suponen un 42 % del documento.

Con el uso de las minimizaciones de forma independiente también se han conseguido buenos resultados (en este caso para el archivo con cinco registros). Por ejemplo, al permitir la omisión de las etiquetas de fin de elemento, se ha consiguido una reducción del 46 % del marcado. Sólo con la eliminación del carácter tagc de algunas etiquetas, que fue una de las propuestas para MicroXML [7], se ha consiguido una reducción del 9 % y al omitir el identificador en las etiquetas de fin, la reducción es de casi un 28 %. Esta última opción parece de las más interesantes por la reducción que consigue, porque no complica la legibilidad, ni tiene dificultad de aplicación en documentos de este tipo, porque no parece complicar el analizador y porque no obligaría a la inclusión de una DTD.

Está claro que a la hora de mantener o no características opcionales de este tipo en un lenguaje de marcado no pueden tenerse en cuenta sólo los argumentos en cuanto al tamaño de los documentos, pero lo que sí puede concluirse con estos resultados, es que el ahorro de memoria no sería de unos pocos bytes, sino que, en este caso, estamos hablando de reducir documentos a menos de la mitad de su tamaño original.

Como posibles líneas de trabajo futuras, sería interesante hacer un estudio más exhaustivo de la característica SHORTREF en bases de datos con un número fijo de campos. Para

estos casos, las reducciones de documentos conseguidas serían aún mayores.

Por otro lado, podrían desarrollarse programas que se encarguen de minimizar los archivos, lo que ocultaría la posible complejidad que supone la aplicación de minimizaciones al usuario.

Bibliografía

[1] BRYAN, Martin. *Web SGML and HTML 4.0 Explained* . Addison Wesley Longman, 1997. Disponible en Internet: <http://www.isthought.co.uk/book/home.htm#contents>.

[2] CLARK, James. *Comparison of SGML and XML.*1997. Disponible en Internet: <http://www.w3.org/TR/NOTE-sgml-xml-971215>.

[3] CLARK, James. *James Clark's MicroXML parser.* 2012. Disponible en Internet: <https://github.com/jclark/microxml-js>.

[4] CLARK, James. *James Clark's SGML parser SP.* Disponible en Internet: *<http://xml.coverpages.org/sp-ann.html>*.

[5] CLARK, James. *MicroXML* .13 Diciembre 2010. Disponible en Internet: <http://blog.jclark.com/2010/12/microxml.html>.

[6] CLARK, James y COWAN, John. *MicroXML* Specification. 1 Octubre 2012. Disponible en Internet: <https://dvcs.w3.org/hg/microxml/raw-file/tip/spec/microxml.html>.

[7] CLARK, James. *More on MicroXML* .18 Diciembre 2010. Disponible en Internet: <http://blog.jclark.com/2010/12/more-on-microxml.html>.

[8] CLARK, James. *SP. An SGML System Conforming to International Standard ISO 8879 – Standard Generalized Markup Language* [en línea]. Disponible en Internet: <http://www.jclark.com/sp/index.htm>.

[9] COVERPAGES. Public SGML/XML Software. Disponible en Internet: <http://xml.coverpages.org/publicSW.html>.

[10] COWAN, John. *MicroXML. The MicroLark parser.* 21 Enero 2011. Disponible en Internet: <http://recycledknowledge.blogspot.com.es/2011/01/microlark-parser.html>.

[11] COWAN, John. *MicroXML. Editor's Draft.*30 Junio 2011. Disponible en Internet: <http://www.ccil.org/~cowan/MicroXML.html>.

[12] COWAN, John. *MicroXML: Who, What, When, Where, Why.* 2012. Disponible en Internet: <http://www.w3.org/community/microxml/wiki/images/b/b4/Microxml-2012.pdf>.

[13] COWAN, John. *XML A Basic Overview.* 1998. Disponible en Internet:<http://ccil.org/~cowan/>.

[14] DEROSE, Steven J. *The SGML Faq Book. Understanding the Foundation of HTML and XML.* Kluwer Academic Publishers, 1997. ISBN-10: 0792399439 .

[15] GOLDFARB, Charles F. *The SGML Handbook.* Edited and with a foreword by Yuri Rubinsky. Oxford University Press, 1990. ISBN 0198537379.

[16] GREEN, Stephen D. *MicroXML and MicroXSD.* 6 Marzo 2011. Disponible en Internet: <http://www.stephengreenxml.org.uk/MicroXML%20and%20MicroXSD.pdf>.

[17] GROSS, Mike. *Conversion to XML: Should I stay with SGML?.* Disponible en Internet: <http://www.dclab.com/sgmltoxml.asp>.

[18] International SGML Users' Group. *SGML Users' Group History . 1990.* Disponible en Internet: <http://xml.coverpages.org/sgmlhist0.html>.

[19] ISO 8879: 1986(E). *Information processing - Text and office systems - Standard Generalized Markup Language (SGML).*

[20] KIMBER, E. *Why I Want the SGML LINK Feature.*1996. Disponible en Internet: <http://xml.coverpages.org/kimber1-link95.html>.

[21] LAMARCA, María Jesús. *Hipertexto: El nuevo concepto de documento en la cultura de la imagen* . Tesis doctoral. Universidad Complutense de Madrid. Facultad de Ciencias de la Información. Dpto. de Biblioteconomía y Documentación. Disponible en Internet: <http://www.hipertexto.info>.

[22] LA QUEY, Robert E. *SML: Simplifying XML.* 24 Noviembre 1999. Disponible en Internet: <http://www.xml.com/pub/a/1999/11/sml/>.

[23] MALER, Eve y EL ANDALOUSSI, Jeanne. *Developing SGML DTDs: From Text to Model to Markup*, [en línea]. [New Jersey]: Prentice Hall, 1995, 10/02/2008. Disponible en Internet: <http://www.xmlgrrl.com/publications/DSDTD/>. ISBN 0133098818.

[24] MicroXML Community Group. Disponible en Internet: <http://www.w3.org/community/microxml/>.

[25] MicroXML Community Group. *Implementations.* Disponible en Internet: <http://www.w3.org/community/microxml/wiki/Implementations>.

[26] MicroXML Community Group. *Research.* Disponible en Internet: <http://www.w3.org/community/microxml/wiki/Main_Page>.

[27] NAGGUM, Erik. *SGML: Erik Naggum's Brief Description.* 7 Febrero 1995. Artículo: 7689 de comp.text.sgml . Disponible en Internet: <http://xml.coverpages.org/naggumWhat.html>.

[28] NAGGUM, Erik. *Arguments against SGML.* Disponible en Internet: <http://xml.coverpages.org/naggumAgainstSGML.html>.

[29] OGBUJI, Uche. *Introducing MicroXML, Part 1: Explore the basic principles of MicroXML* . Disponible en Internet: <http://www.ibm.com/developerworks/library/x-microxml1/index.html>.

[30] OGBUJI, Uche. *Introducing MicroXML, Part 2: Process MicroXML with MicroLark*. Disponible en Internet: <http://www.ibm.com/developerworks/library/x-microxml2/index.html>.

[31] OGBUJI, Uche. *Introducing MicroXML. XML Prague*. 10 Febrero 2013. Disponible en Internet: <http://archive.xmlprague.cz/2013/presentations/ Introducing_MicroXML.pdf>.

[32] OMNIMARK. *Guide to OmniMark 10.1.0*. Omnimark no soporta Datatag. Disponible en Internet: <http://developers.omnimark.com/docs/html/error/0024.htm>.

[33] OPENJADE. OpenJade Distribution Page. Disponible en Internet: <http://openjade.sourceforge.net/>

[34] T*e*xt Encoding Initiative. *A Gentle Introduction to SGML*.Disponible en Internet: <http://www.isgmlug.org/sgmlhelp/g-index.htm>.

[35] T*e*xt Encoding Initiative. *A Gentle Introduction to XML*. Disponible en Internet: <http://www.tei-c.org/release/doc/tei-p4-doc/html/SG.html>.

[36] WALSH, Norman. *A Technical Introduction to XML*. 1998. Disponible en Internet: <http://nwalsh.com/docs/articles/xml/>.

[37] WALSH, Norman. *Converting an SGML DTD to XML*. 1998. Disponible en Internet: <http://www.xml.com/lpt/a/288>.

[38] WALSH, Norman. *XML v.next* . 4 Abril 2011. Disponible en Internet: <http://norman.walsh.name/2011/03/28/XMLvNext>.

[39] WIEDERHOLD, Gio. *Movies Database Documentation*. Disponible en Internet: <http://infolab.stanford.edu/pub/movies/doc.html >.

[40] WOHLER, Wayne. *SGML Declarations*.1994. Disponible en Internet: <http://xml.coverpages.org/wlw11.html>.

[41] W3C. *Extensible Markup Language (XML) 1.0, Fifth Edition*. 2008. Disponible en Internet: <http://www.w3.org/TR/REC-xml/>.

[42] W3C. Namespaces in XML. Disponible en Internet:<http://www.w3.org/TR/1999/REC-xml-names-19990114/>.

[43] W3C España. *Guía Breve de XHTML* . 2008.Disponible en Internet: <http://www.w3c.es/divulgacion/guiasbreves/XHTML>.

[44] W3C España. *Preguntas frecuentes sobre HTML y XHTML*.2004. Disponible en Internet: <http://www.w3c.es/Traducciones/es/Markup/2004/xhtml-faq.htm#need>.

[45] W3C. *HTML 4.01 Specification.*1999. Disponible en Internet: <http://www.w3.org/TR/html401/>.

[46] W3C Public Mailing List Archives. Tim Bray. *Ejemplo de la discusión a favor de mantener las minimizaciones de la lista de atributos en XML.*1996. Disponible en Internet: <http://lists.w3.org/Archives/Public/w3c-sgml-wg/1996Sep/0048.html>.

[47] W3C Public Mailing List Archives. Arjun Ray. *Ejemplo de la discusión a favor de mantener la posibilidad de omitir el identificador en las etiquetas de fin en XML.*1996. Disponible en Internet:<http://lists.w3.org/Archives/Public/w3c-sgml-wg/1996Sep/0143.html>.

[48] W3C Markup Validator. Instalación del analizador. Disponible en Internet: <http://validator.w3.org/docs/install.html#install-prereq-sp>.

[49] XHTML2 Working Group Home Page. *Mission of the XHTML2 Working Group.* 2010. Disponible en Internet: <http://www.w3.org/MarkUp/>.

Apéndice A

Aplicación de las minimizaciones a cincuenta registros de la base de datos

En este anexo se incluye el documento SGML con el que se trabaja en este proyecto y que contiene cincuenta registros de una base de datos. Se incluye, en primer lugar, el documento completo y, a continuación, el documento obtenido al aplicar las técnicas de minimización que se han considerado adecuadas para el mismo.

A.1. Documento completo

```
<!DOCTYPE movies SYSTEM "dtd_modificada.dtd">
<movies>
 <film fid="H1">
  <t>Always Tell Your Wife</t>
  <year>1922</year>
  <dirs>
   <dir dirk="R">
    <dirn>Se.Hicks</dirn>
   </dir>
   <dir dirk="R">
    <dirn>Hitchcock</dirn>
   </dir>
  </dirs>
  <prods>
   <prod prodk="R">
    <pname>Lasky</pname>
   </prod>
  </prods>
  <studios>
   <studio>Famous</studio>
  </studios>
```

```
 <prcs>
  <prc>sbw</prc>
 </prcs>
 <cats>
  <cat>Dram</cat>
 </cats>
</film>

<film fid="H2">
 <t>Number Thirteen</t>
 <year>1922</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Hitchcock</pname>
  </prod>
 </prods>
 <studios>
  <studio>Islington</studio>
  <distributor>Famous</distributor>
 </studios>
 <prcs>
  <prc>sbw</prc>
  <prctext>unfinished</prctext>
 </prcs>
</film>

<film fid="H3">
 <t>Woman to Woman</t>
 <year>1922</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
  <dir dirk="N">
   <dirn>Cutts</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Balcon</pname>
  </prod>
 </prods>
```

```
 <studios>
  <studio>B-S-F</studio>
  <distributor>Wardour</distributor>
 </studios>
 <prcs>
  <prc>sbw</prc>
 </prcs>
 <cats>
  <cat>Dram</cat>
 </cats>
 <loc>
  <site>
   <siteplace>GB</siteplace>
  </site>
 </loc>
 <error>same(GCt27), y(1926)</error>
</film>

<film fid="H4">
 <t>The Passionate Adventure</t>
 <year>1924</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
  <dir dirk="N">
   <dirn>Cutts</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Balcon</pname>
  </prod>
 </prods>
 <studios>
  <studio>Gainsborough</studio>
  <distributor>GaumontD</distributor>
 </studios>
 <prcs>
  <prc>sbw</prc>
 </prcs>
</film>

<film fid="H5">
 <t>The Blackguard</t>
 <year>1925</year>
 <dirs>
```

```
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
  <dir dirk="N">
   <dirn>Cutts</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Balcon</pname>
  </prod>
 </prods>
 <studios>
  <studio>UFA</studio>
  <distributor>Wardour</distributor>
 </studios>
 <prcs>
  <prc>sbw</prc>
 </prcs>
</film>

<film fid="H6">
 <t>The Pleasure Garden</t>
 <year> 1925  </year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Balcon</pname>
  </prod>
 </prods>
 <studios>
  <studio>Gainsborough and Emelka</studio>
  <distributor>Wardour</distributor>
 </studios>
 <prcs>
  <prc>sbw</prc>
 </prcs>
 <cats>
  <cat>Dram</cat>
 </cats>
 <awards>
  <aw>
   <awtype>H</awtype>
```

```
    <awattr>0</awattr>
   </aw>
  </awards>
  <loc>
   <site>
    <sitename>Pleasure Garden</sitename>
    <siteclass>theater</siteclass>
   </site>
  </loc>
  <people>
   <authors>
    <names>
     <name>Oliver Sandys</name>
    </names>
   </authors>
   <writers>
    <names>
     <name>Eliot Stannard</name>
    </names>
   </writers>
   <cingraphs>
    <names>
     <kname>Ventimiglia</kname>
    </names>
   </cingraphs>
  </people>
 </film>

 <film fid="H7">
  <t>The Mountain Eagle</t>
  <alts>
   <alt>
    <altt>Fear O'God</altt>
    <altwhy>USA</altwhy>
   </alt>
  </alts>
  <year>1926</year>
  <dirs>
   <dir dirk="R">
    <dirn>Hitchcock</dirn>
   </dir>
  </dirs>
  <prods>
   <prod prodk="R">
    <pname>Balcon</pname>
   </prod>
  </prods>
```

```
 <studios>
  <studio>Gainsborough</studio>
  <studio>Emelka</studio>
  <distributor>Wardour</distributor>
 </studios>
 <prcs>
  <prc>sbw</prc>
  <prctext>no prints left</prctext>
 </prcs>
 <cats>
  <cat>Dram</cat>
 </cats>
 <awards>
  <aw>
   <awtype>H</awtype>
   <awattr>0</awattr>
  </aw>
 </awards>
 <people>
  <writers>
   <names>
    <name>Eliot Stannard</name>
   </names>
  </writers>
  <cingraphs>
   <names>
    <kname>Ventimiglia</kname>
   </names>
  </cingraphs>
 </people>
</film>

<film fid="H8">
 <t>The Lodger: A Story of The London Fog</t>
 <alts>
  <alt>
   <altt>The Case of Jonathan Drew</altt>
   <altwhy>USA</altwhy>
  </alt>
 </alts>
 <year>1926</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
```

```
 <prod prodk="R">
  <pname>Balcon</pname>
 </prod>
</prods>
<studios>
 <studio>Gainsborough</studio>
 <distributor>Wardour</distributor>
</studios>
<prcs>
 <prc>sbw</prc>
</prcs>
<cats>
 <cat>Susp</cat>
</cats>
<loc>
 <site>
  <sitename>London</sitename>
  <siteclass>city</siteclass>
  <siteplace>GB</siteplace>
 </site>
</loc>
<people>
 <authors>
  <names>
   <name>Mary Belloc Lowndes</name>
  </names>
  <bt>The Lodger</bt>
 </authors>
 <writers>
  <names>
   <name>Hitchcock</name>
   <name>Eliot Stannard</name>
  </names>
 </writers>
 <cingraphs>
  <names>
   <kname>Ventimiglia</kname>
  </names>
 </cingraphs>
</people>
<notes>
 <money>
  <cost>12K</cost>
  <moneynotes>in UK lbs</moneynotes>
 </money>
 <facts>Topic:Jack the Ripper</facts>
</notes>
```

```
</film>

<film fid="H9">
 <t>Downhill</t>
 <alts>
  <alt>
   <altt>When Boys Leave Home</altt>
  </alt>
 </alts>
 <year>1927</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Balcon</pname>
  </prod>
 </prods>
 <studios>
  <studio>Islington; Gainsborough</studio>
  <distributor>Wardour</distributor>
 </studios>
 <prcs>
  <prc>sbw</prc>
 </prcs>
 <cats>
  <cat>Susp</cat>
 </cats>
 <people>
  <authors>
   <names>
    <name>Constance Collier</name>
   </names>
  </authors>
  <writers>
   <names>
    <name>Eliot Stannard</name>
   </names>
  </writers>
  <cingraphs>
   <names>
    <kname>Ventimiglia</kname>
   </names>
  </cingraphs>
 </people>
```

```
 <error>W(Ivor Novello)</error>
</film>

<film fid="H10">
 <t>Easy Virtue</t>
 <year>1927</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Balcon</pname>
  </prod>
 </prods>
 <studios>
  <studio>Gainsborough</studio>
  <distributor>Wardour</distributor>
 </studios>
 <prcs>
  <prc>sbw</prc>
 </prcs>
 <cats>
  <cat>Susp</cat>
 </cats>
 <loc>
  <site>
   <sitedes>artist's studio</sitedes>
   <siteclass>work</siteclass>
   <siteplace>England</siteplace>
  </site>
  <site>
   <siteclass>hotel</siteclass>
   <siteat>Riviera</siteat>
   <siteplace>France</siteplace>
  </site>
  <site>
   <sitedes>country mansion</sitedes>
   <siteclass>estate</siteclass>
   <siteplace>England</siteplace>
  </site>
 </loc>
 <people>
  <authors>
   <names>
    <name>Noel Coward</name>
```

```
    </names>
   </authors>
   <writers>
    <names>
     <name>Eliot Stannard</name>
    </names>
   </writers>
   <cingraphs>
    <names>
     <name>Claude McDonnell</name>
    </names>
   </cingraphs>
  </people>
  <notes>
   <source>
    <seen>11Feb2000</seen>
   </source>
  </notes>
 </film>

 <film fid="H11">
  <t>The Ring</t>
  <year>1927</year>
  <dirs>
   <dir dirk="R">
    <dirn>Hitchcock</dirn>
   </dir>
  </dirs>
  <prods>
   <prod prodk="R">
    <pname>J.Maxwell</pname>
   </prod>
  </prods>
  <studios>
   <studio>BIP Elstree</studio>
   <distributor>Wardour</distributor>
  </studios>
  <prcs>
   <prc>sbw</prc>
  </prcs>
  <cats>
   <cat>Susp</cat>
  </cats>
  <loc>
   <site>
    <sitedes>boxing rings</sitedes>
    <siteclass>sports</siteclass>
```

```
    <siteplace>England</siteplace>
   </site>
   <site>
    <sitename>Albert Hall</sitename>
    <siteclass>theater</siteclass>
    <siteat>London</siteat>
    <siteplace>GB</siteplace>
   </site>
  </loc>
  <people>
   <writers>
    <names>
     <name>Hitchcock</name>
     <name>Alma Reveille</name>
    </names>
   </writers>
   <cingraphs>
    <names>
     <name>Claude McDonnell</name>
    </names>
   </cingraphs>
  </people>
  <notes>
   <source>
    <seen>11Feb2000</seen>
   </source>
  </notes>
</film>

<film fid="H12">
 <t>The Farmer's Wife</t>
 <year>1928</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>J.Maxwell</pname>
  </prod>
 </prods>
 <studios>
  <studio>BIP Elstree</studio>
  <distributor>Wardour</distributor>
 </studios>
 <prcs>
```

```
  <prc>sbw</prc>
 </prcs>
 <cats>
  <cat>Susp</cat>
 </cats>
 <people>
  <authors>
   <names>
    <name>Eden Philpotts</name>
   </names>
  </authors>
  <writers>
   <names>
    <name>Hitchcock</name>
   </names>
  </writers>
  <cingraphs>
   <names>
    <name>J.Cox</name>
   </names>
  </cingraphs>
 </people>
</film>

<film fid="H13">
 <t>Champagne</t>
 <year>1928</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>J.Maxwell</pname>
  </prod>
 </prods>
 <studios>
  <studio>BIP Elstree</studio>
  <distributor>Wardour</distributor>
 </studios>
 <prcs>
  <prc>sbw</prc>
 </prcs>
 <cats>
  <cat>Romt</cat>
 </cats>
```

```
 <people>
  <authors>
   <names>
    <name>Walter Mycroft</name>
   </names>
  </authors>
  <writers>
   <names>
    <name>Eliot Stannard</name>
   </names>
  </writers>
  <cingraphs>
   <names>
    <name>J.Cox</name>
   </names>
  </cingraphs>
 </people>
</film>

<film fid="H14">
 <t>Harmony Heaven</t>
 <year>1929</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
  <dir dirk="N">
   <dirn>Thomas Bentley</dirn>
  </dir>
 </dirs>
 <studios>
  <studio>BIP</studio>
  <distributor>France-Societ\'e des Cin\'e-Romans</distributor>
 </studios>
 <prcs>
  <prc>col</prc>
 </prcs>
 <cats>
  <cat>Susp</cat>
 </cats>
 <awards>
  <aw>
   <awtype>H</awtype>
   <awattr>0</awattr>
  </aw>
 </awards>
 <people>
```

```
  <writers>
   <names>
    <name>Arthur Wimperis</name>
    <name>Randall Faye</name>
   </names>
  </writers>
  <cingraphs>
   <names>
    <name>J.Cox</name>
   </names>
  </cingraphs>
  <composers>
   <names>
    <name>Eddie Pola</name>
    <name>Edward Brandt</name>
   </names>
  </composers>
 </people>
</film>

<film fid="H15">
 <t>The Manxman</t>
 <year>1929</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>J.Maxwell</pname>
  </prod>
 </prods>
 <studios>
  <studio>BIP Elstree</studio>
  <distributor>Wardour</distributor>
 </studios>
 <prcs>
  <prc>sbw</prc>
 </prcs>
 <cats>
  <cat>Susp</cat>
 </cats>
 <awards>
  <aw>
   <awtype>H</awtype>
   <awattr>0</awattr>
```

```
  </aw>
 </awards>
 <loc>
  <site>
   <siteclass>csd</siteclass>
   <siteat>Scotland</siteat>
   <siteplace>GB</siteplace>
  </site>
 </loc>
 <people>
  <authors>
   <names>
    <name>Hall Caine</name>
   </names>
  </authors>
  <writers>
   <names>
    <name>Eliot Stannard</name>
   </names>
  </writers>
  <cingraphs>
   <names>
    <name>J.Cox</name>
   </names>
  </cingraphs>
 </people>
</film>

<film fid="H16">
 <t>Blackmail</t>
 <year>1929</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>J.Maxwell</pname>
  </prod>
 </prods>
 <studios>
  <studio>BIP Elstree</studio>
  <distributor>Wardour</distributor>
 </studios>
 <prcs>
  <prc>bnw</prc>
```

```
  <prctext>first sound, added</prctext>
 </prcs>
 <cats>
  <cat>Susp</cat>
 </cats>
 <loc>
  <site>
   <sitename>British Museum</sitename>
   <siteclass>museum</siteclass>
   <siteat>London</siteat>
   <siteplace>GB</siteplace>
  </site>
 </loc>
 <people>
  <writers>
   <names>
    <name>Hitchcock</name>
    <name>Benn W. Levy</name>
    <name>Charles Bennett</name>
   </names>
  </writers>
  <cingraphs>
   <names>
    <name>J.Cox</name>
   </names>
  </cingraphs>
 </people>
 <notes>
  <source>
   <seen>9Apr1990</seen>
  </source>
 </notes>
</film>

<film fid="H17">
 <t>Elstree Calling</t>
 <year>1929</year>
 <dirs>
  <dir dirk="R">
   <dirn>Brunel</dirn>
  </dir>
  <dir dirk="N">
   <dirn>Hitchcock</dirn>
  </dir>
  <dir dirk="N">
   <dirn>Andr\'e Charlot</dirn>
  </dir>
```

```
  <dir dirk="N">
   <dirn>Jack Hulbert</dirn>
  </dir>
  <dir dirk="N">
   <dirn>Paul Murray</dirn>
  </dir>
 </dirs>
 <studios>
  <studio>BIP Elstree</studio>
 </studios>
 <prcs>
  <prc>bnw</prc>
 </prcs>
 <cats>
  <cat>Musc</cat>
 </cats>
 <loc>
  <site>
   <sitedes>studio</sitedes>
   <siteclass>work</siteclass>
  </site>
 </loc>
 <people>
  <writers>
   <names>
    <name>Ivor Novello</name>
   </names>
  </writers>
 </people>
 <notes>
  <facts>Brunel was the supervisor</facts>
 </notes>
</film>

<film fid="H18">
 <t>Juno And The Paycock</t>
 <year>1930</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>J.Maxwell</pname>
  </prod>
 </prods>
```

```
 <studios>
  <studio>BIP Elstree</studio>
  <distributor>Wardour</distributor>
 </studios>
 <prcs>
  <prc>bnw</prc>
 </prcs>
 <cats>
  <cat>Susp</cat>
 </cats>
 <awards>
  <aw>
   <awtype>H</awtype>
   <awattr>*</awattr>
  </aw>
 </awards>
 <loc>
  <site>
   <sitename>Dublin</sitename>
   <siteclass>city</siteclass>
   <siteplace>Ireland</siteplace>
  </site>
 </loc>
 <people>
  <authors>
   <names>
    <name>Sean O'Casey</name>
   </names>
  </authors>
  <writers>
   <names>
    <name>Hitchcock</name>
    <name>Alma Reville</name>
   </names>
  </writers>
  <cingraphs>
   <names>
    <name>J.Cox</name>
   </names>
  </cingraphs>
 </people>
</film>

<film fid="H19">
 <t>Murder!</t>
 <year>1930</year>
 <dirs>
```

```
 <dir dirk="R">
  <dirn>Hitchcock</dirn>
 </dir>
</dirs>
<prods>
 <prod prodk="R">
  <pname>J.Maxwell</pname>
 </prod>
 <prod prodk="X">
  <pname>Alfred Abel</pname>
 </prod>
 <prodnote>Abel produced the German version</prodnote>
</prods>
<studios>
 <studio>BIP Elstree</studio>
 <distributor>Wardour</distributor>
</studios>
<prcs>
 <prc>bnw</prc>
</prcs>
<cats>
 <cat>Myst</cat>
</cats>
<awards>
 <aw>
  <awtype>H</awtype>
  <awattr>**</awattr>
 </aw>
</awards>
<loc>
 <site>
  <siteclass>theater</siteclass>
  <siteat>London</siteat>
  <siteplace>GB</siteplace>
 </site>
</loc>
<people>
 <authors>
  <names>
   <name>Clarence Dane</name>
   <name>Helen Simpson</name>
  </names>
  <bt>Enter Sir John</bt>
 </authors>
 <writers>
  <names>
   <kname>Hitchcock</kname>
```

```
    <name>Alma Reville</name>
   </names>
  </writers>
  <visuals>
   <names>
    <name>Winifred Ashton</name>
   </names>
  </visuals>
  <cingraphs>
   <names>
    <name>J.Cox</name>
   </names>
  </cingraphs>
 </people>
 <notes>
  <source>
   <seen>10apr1990</seen>
  </source>
 </notes>
 <error>V</error>
</film>

<film fid="H20">
 <t>The Skin Game</t>
 <year>1931</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>J.Maxwell</pname>
  </prod>
 </prods>
 <studios>
  <studio>BIP Elstree</studio>
  <distributor>Wardour</distributor>
 </studios>
 <prcs>
  <prc>bnw</prc>
 </prcs>
 <cats>
  <cat>Dram</cat>
 </cats>
 <people>
  <authors>
```

```
    <names>
     <name>John Galsworthy</name>
    </names>
   </authors>
   <writers>
    <names>
     <name>Hitchcock</name>
     <name>Alma Reville</name>
    </names>
   </writers>
   <cingraphs>
    <names>
     <name>J.Cox</name>
     <name>Charles Martin</name>
    </names>
   </cingraphs>
  </people>
  <error>C</error>
 </film>

 <film fid="H21">
  <t>Rich and Strange</t>
  <alts>
   <alt>
    <altt>East of Shanghai</altt>
    <altwhy>USA</altwhy>
   </alt>
  </alts>
  <year>1932</year>
  <dirs>
   <dir dirk="R">
    <dirn>Hitchcock</dirn>
   </dir>
  </dirs>
  <prods>
   <prod prodk="R">
    <pname>J.Maxwell</pname>
   </prod>
  </prods>
  <studios>
   <studio>BIP Elstree</studio>
   <distributor>Wardour</distributor>
  </studios>
  <prcs>
   <prc>bnw</prc>
  </prcs>
  <cats>
```

```
 <cat>susp</cat>
</cats>
<loc>
 <site>
  <sitename>London</sitename>
  <siteclass>city</siteclass>
  <siteplace>England</siteplace>
 </site>
 <site>
  <sitename>Paris</sitename>
  <siteclass>city</siteclass>
  <siteplace>France</siteplace>
 </site>
 <site>
  <sitename>Marseille</sitename>
  <siteclass>city</siteclass>
  <siteplace>France</siteplace>
 </site>
 <site>
  <siteplace>Singapore</siteplace>
 </site>
 <site>
  <siteclass>ship</siteclass>
  <siteat>sea</siteat>
  <siteplace>Pacific</siteplace>
 </site>
</loc>
<people>
 <authors>
  <names>
   <name>Dale Collins</name>
  </names>
 </authors>
 <writers>
  <names>
   <name>Alma Reville</name>
   <name>Val Valentine</name>
  </names>
 </writers>
 <cingraphs>
  <names>
   <name>J.Cox</name>
   <name>Charles Martin</name>
  </names>
 </cingraphs>
</people>
<notes>
```

```
  <source>
   <seen>16apr1990</seen>
  </source>
 </notes>
 <error>C</error>
</film>

<film fid="H22">
 <t>Number Seventeen</t>
 <year>1932</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>J.Maxwell</pname>
  </prod>
 </prods>
 <studios>
  <studio>BIP Elstree</studio>
  <distributor>Wardour</distributor>
 </studios>
 <prcs>
  <prc>bnw</prc>
 </prcs>
 <cats>
  <cat>Myst</cat>
 </cats>
 <awards>
  <aw>
   <awtype>H</awtype>
   <awattr>*</awattr>
  </aw>
 </awards>
 <loc>
  <site>
   <sitename>London</sitename>
   <siteclass>city</siteclass>
   <siteplace>England</siteplace>
  </site>
  <site>
   <siteclass>train</siteclass>
   <siteplace>England</siteplace>
  </site>
  <site>
```

```
   <sitename>Dover</sitename>
   <siteclass>town</siteclass>
   <siteplace>England</siteplace>
  </site>
 </loc>
 <people>
  <authors>
   <names>
    <name>Jefferson Farejon</name>
   </names>
  </authors>
  <writers>
   <names>
    <name>Jefferson Farejon</name>
    <name>Rodney Auckland</name>
   </names>
  </writers>
  <cingraphs>
   <names>
    <name>J.Cox</name>
   </names>
  </cingraphs>
 </people>
 <notes>
  <source>
   <seen>16apr1990</seen>
  </source>
 </notes>
 <error>W</error>
</film>

<film fid="H23">
 <t>Lord Camber's Ladies</t>
 <year>1932</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcok</dirn>
  </dir>
  <dir dirk="N">
   <dirn>Benn~Levy</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Hitchcock</pname>
  </prod>
 </prods>
```

```
 <studios>
  <studio>BIP Elstree</studio>
  <distributor>Wardour</distributor>
 </studios>
 <prcs>
  <prc>bnw</prc>
 </prcs>
 <cats>
  <cat>Susp</cat>
 </cats>
 <loc>
  <site>
   <sitedes>castle</sitedes>
   <siteclass>palace</siteclass>
   <siteplace>GB</siteplace>
  </site>
 </loc>
 <people>
  <authors>
   <names>
    <name>H.A. Vachell</name>
   </names>
   <bt>The Case of Lady Camber</bt>
  </authors>
  <writers>
   <names>
    <name>Hitchcock</name>
    <name>Jefferson Farejon</name>
   </names>
  </writers>
 </people>
</film>

<film fid="H24">
 <t>Waltzes From Vienna</t>
 <alts>
  <alt>
   <altt>Strauss's Great Waltz</altt>
   <altwhy>USA</altwhy>
  </alt>
 </alts>
 <year>1933</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
```

```
<prods>
 <prod prodk="N">
  <pname>Tom Arnold</pname>
 </prod>
</prods>
<studios>
 <studio>Lime Grove</studio>
 <distributor>G.F.D.</distributor>
</studios>
<prcs>
 <prc>bnw</prc>
</prcs>
<cats>
 <cat>Romt</cat>
 <cat>Comd</cat>
</cats>
<awards>
 <aw>
  <awtype>H</awtype>
  <awattr>0</awattr>
 </aw>
 <aw>
  <awtype>VIP</awtype>
  <awattr>least favorite of Hitchcock</awattr>
 </aw>
</awards>
<loc>
 <site>
  <sitename>Vienna</sitename>
  <siteclass>city</siteclass>
  <siteplace>Austria</siteplace>
 </site>
</loc>
<people>
 <writers>
  <names>
   <name>Alma Reveille</name>
   <name>Bolton</name>
  </names>
 </writers>
 <composers>
  <names>
   <name>Johann Strauss~sr.</name>
   <name>Johann Strauss~jr.</name>
  </names>
 </composers>
</people>
```

```
</film>

<film fid="H25">
 <t>The Man Who Knew Too Much</t>
 <year>1934</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Balcon</pname>
  </prod>
  <prod prodk="X">
   <pname>Montagu</pname>
  </prod>
 </prods>
 <studios>
  <studio>Gaumont Lime Grove</studio>
  <distributor>G.F.D.</distributor>
 </studios>
 <prcs>
  <prc>bnw</prc>
 </prcs>
 <cats>
  <cat>Susp</cat>
 </cats>
 <awards>
  <aw>
   <awtype>H</awtype>
   <awattr>***</awattr>
  </aw>
 </awards>
 <loc>
  <site>
   <sitename>St.Moritz</sitename>
   <siteclass>mountains</siteclass>
   <siteplace>Switzerland</siteplace>
  </site>
  <site>
   <sitename>London</sitename>
   <siteclass>city</siteclass>
   <siteplace>GB</siteplace>
  </site>
 </loc>
 <people>
```

```
  <authors>
   <names>
    <name>D.B. Wyndham-Lewis</name>
    <name>Edwin Greenwood</name>
   </names>
  </authors>
  <writers>
   <names>
    <name>A.R.Rawlinson</name>
    <name>Charles Bennett</name>
    <name>D.B. Wyndham-Lewis</name>
    <name>Edwin Greenwood</name>
   </names>
  </writers>
  <cingraphs>
   <names>
    <name>Curt Courant</name>
   </names>
  </cingraphs>
  <composers>
   <names>
    <kname>Louis~Levy</kname>
    <name>Arthur Benjamin</name>
   </names>
  </composers>
 </people>
</film>

<film fid="H26">
 <t>The 39 Steps</t>
 <year>1935</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Balcon</pname>
  </prod>
  <prod prodk="X">
   <pname>Montagu</pname>
  </prod>
 </prods>
 <studios>
  <studio>Gaumont Lime Grove</studio>
  <distributor>G.F.D.</distributor>
```

```
 </studios>
 <prcs>
  <prc>bnw</prc>
 </prcs>
 <cats>
  <cat>Susp</cat>
 </cats>
 <awards>
  <aw>
   <awtype>H</awtype>
   <awattr>****</awattr>
  </aw>
 </awards>
 <loc>
  <site>
   <siteclass>theater</siteclass>
   <siteat>London</siteat>
   <siteplace>GB</siteplace>
  </site>
 </loc>
 <people>
  <authors>
   <names>
    <name>John Buchan</name>
   </names>
  </authors>
  <writers>
   <names>
    <name>Charles Bennett</name>
    <name>Alma Reveille</name>
   </names>
  </writers>
  <cingraphs>
   <names>
    <name>Bernard Knowles</name>
   </names>
  </cingraphs>
  <composers>
   <names>
    <name>Louis Levy</name>
   </names>
  </composers>
 </people>
</film>

<film fid="H27">
 <t>Secret Agent</t>
```

```
<year>1936</year>
<dirs>
 <dir dirk="R">
  <dirn>Hitchcock</dirn>
 </dir>
</dirs>
<prods>
 <prod prodk="R">
  <pname>Balcon</pname>
 </prod>
 <prod prodk="X">
  <pname>Montagu</pname>
 </prod>
</prods>
<studios>
 <studio>Gaumont Lime Grove</studio>
 <distributor>G.F.D.</distributor>
</studios>
<prcs>
 <prc>bnw</prc>
</prcs>
<cats>
 <cat>Susp</cat>
</cats>
<loc>
 <site>
  <sitename>London</sitename>
  <siteclass>city</siteclass>
  <siteplace>England</siteplace>
 </site>
 <site>
  <sitedes>chocolate factory</sitedes>
  <siteclass>work</siteclass>
  <siteplace>Switzerland</siteplace>
 </site>
</loc>
<people>
 <authors>
  <names>
   <name>S.Maugham</name>
  </names>
  <bt>Ashenden</bt>
 </authors>
 <writers>
  <names>
   <name>Charles Bennett</name>
   <name>Campbell Dixon</name>
```

```
    </names>
   </writers>
   <cingraphs>
    <names>
     <name>Bernard Knowles</name>
    </names>
   </cingraphs>
  </people>
  <error>W</error>
 </film>

 <film fid="H28">
  <t>Sabotage</t>
  <alts>
   <alt>
    <altt>The Woman Alone</altt>
    <altwhy>USA</altwhy>
   </alt>
  </alts>
  <year>1936</year>
  <dirs>
   <dir dirk="R">
    <dirn>Hitchcock</dirn>
   </dir>
  </dirs>
  <prods>
   <prod prodk="R">
    <pname>Balcon</pname>
   </prod>
   <prod prodk="X">
    <pname>Montague</pname>
   </prod>
  </prods>
  <studios>
   <studio>Lime Grove</studio>
   <distributor>G.F.D.</distributor>
  </studios>
  <prcs>
   <prc>bnw</prc>
  </prcs>
  <cats>
   <cat>Susp</cat>
  </cats>
  <awards>
   <aw>
    <awtype>H</awtype>
    <awattr>***</awattr>
```

```
  </aw>
 </awards>
 <loc>
  <site>
   <siteclass>theater</siteclass>
   <siteat>London</siteat>
   <siteplace>GB</siteplace>
  </site>
 </loc>
 <people>
  <authors>
   <names>
    <name>Conrad</name>
   </names>
   <bt>The Secret Agent</bt>
  </authors>
  <writers>
   <names>
    <name>Charles Bennett</name>
   </names>
  </writers>
  <cingraphs>
   <names>
    <name>Bernard Knowles</name>
   </names>
  </cingraphs>
 </people>
</film>

<film fid="H29">
 <t>Young and Innocent</t>
 <alts>
  <alt>
   <altt>The Girl Was Young</altt>
   <altwhy>USA</altwhy>
  </alt>
 </alts>
 <year>1938</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Black</pname>
  </prod>
```

```
 </prods>
 <studios>
  <studio>Lime Grove and Pinewood</studio>
  <distributor>G.F.D.</distributor>
 </studios>
 <prcs>
  <prc>bnw</prc>
 </prcs>
 <cats>
  <cat>Susp</cat>
 </cats>
 <people>
  <authors>
   <names>
    <name>Josephine Tey</name>
   </names>
  </authors>
  <writers>
   <names>
    <name>Charles Bennett</name>
    <name>Alma Reveille</name>
   </names>
  </writers>
  <cingraphs>
   <names>
    <name>Bernard Knowles</name>
   </names>
  </cingraphs>
 </people>
</film>

<film fid="H30">
 <t>The Lady Vanishes</t>
 <year>1938</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Black</pname>
  </prod>
 </prods>
 <studios>
  <studio>Lime Grove Gainsborough</studio>
  <distributor>MGM</distributor>
```

```
 </studios>
 <prcs>
  <prc>bnw</prc>
 </prcs>
 <cats>
  <cat>Susp</cat>
 </cats>
 <loc>
  <site>
   <siteclass>train</siteclass>
  </site>
  <site>
   <siteplace>Austria</siteplace>
  </site>
 </loc>
 <people>
  <authors>
   <names>
    <name>Ethel Lina White</name>
   </names>
   <bt>‘The Wheel Spins’</bt>
  </authors>
  <writers>
   <names>
    <name>Gilliat</name>
    <name>Launder</name>
   </names>
  </writers>
  <cingraphs>
   <names>
    <name>J.Cox</name>
   </names>
  </cingraphs>
 </people>
 <notes>
  <source>
   <seen>11Jul1988</seen>
  </source>
 </notes>
</film>

<film fid="H31">
 <t>Jamaica Inn</t>
 <year>1939</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
```

```
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Pommer</pname>
  </prod>
  <prod prodk="X">
   <pname>Laughton</pname>
  </prod>
 </prods>
 <studios>
  <studio>Elstree</studio>
  <distributor>Associated British</distributor>
  <distributor>Paramount</distributor>
 </studios>
 <prcs>
  <prc>bnw</prc>
 </prcs>
 <cats>
  <cat>Dram</cat>
 </cats>
 <awards>
  <aw>
   <awtype>W50</awtype>
  </aw>
 </awards>
 <loc>
  <site>
   <sitedes>seashore</sitedes>
   <siteclass>resort</siteclass>
   <siteat>Cornwall</siteat>
   <siteplace>GB</siteplace>
  </site>
 </loc>
 <people>
  <cingraphs>
   <names>
    <name>Bernard Knowles</name>
    <name>Harry Stradling</name>
   </names>
  </cingraphs>
 </people>
 <error>C</error>
</film>

<film fid="H32">
 <t>Rebecca</t>
```

```
<year>1940</year>
<dirs>
 <dir dirk="R">
  <dirn>Hitchcock</dirn>
 </dir>
</dirs>
<prods>
 <prod prodk="R">
  <pname>Selznick</pname>
 </prod>
</prods>
<studios>
 <studio>Selznick</studio>
 <distributor>U.A.</distributor>
</studios>
<prcs>
 <prc>bnw</prc>
</prcs>
<cats>
 <cat>Dram</cat>
</cats>
<awards>
 <aw>
  <awtype>AA</awtype>
 </aw>
 <aw>
  <awtype>AANdir</awtype>
 </aw>
 <aw>
  <awtype>H</awtype>
  <awattr>****</awattr>
 </aw>
</awards>
<loc>
 <site>
  <sitename>Mandalay</sitename>
  <sitedes>castle</sitedes>
  <siteclass>estate</siteclass>
  <siteplace>GB</siteplace>
 </site>
</loc>
<people>
 <authors>
  <names>
   <name>Daphne duMaurier</name>
  </names>
 </authors>
```

```
  <writers>
   <names>
    <kname>J.Harrison</kname>
    <name>Robert E. Sherwood</name>
   </names>
   <pawards>
    <paw>AAN</paw>
   </pawards>
  </writers>
  <visuals>
   <names>
    <kname>Lyle~Wheeler</kname>
   </names>
  </visuals>
  <cingraphs>
   <names>
    <kname>George~Barnes</kname>
   </names>
   <pawards>
    <paw>AA</paw>
   </pawards>
  </cingraphs>
  <composers>
   <names>
    <name>Waxman</name>
   </names>
   <pawards>
    <paw>AAN</paw>
   </pawards>
  </composers>
 </people>
 <notes>
  <source>
   <seen>1989, 25Jan2000</seen>
   <vt>N-HT5</vt>
  </source>
 </notes>
</film>

<film fid="H33">
 <t>Foreign Correspondent</t>
 <year>1940</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
```

```
<prods>
 <prod prodk="R">
  <pname>Wanger</pname>
 </prod>
</prods>
<studios>
 <studio>U.A.</studio>
</studios>
<prcs>
 <prc>bnw</prc>
</prcs>
<cats>
 <cat>Susp</cat>
</cats>
<awards>
 <aw>
  <awtype>H</awtype>
  <awattr>****</awattr>
 </aw>
 <aw>
  <awtype>AAN</awtype>
 </aw>
</awards>
<loc>
 <site>
  <sitename>London</sitename>
  <siteclass>city</siteclass>
  <siteplace>England</siteplace>
 </site>
 <site>
  <siteplace>Netherlands</siteplace>
 </site>
 <site>
  <sitedes>flying boat</sitedes>
  <siteclass>airplane</siteclass>
  <siteplace>Atlantic</siteplace>
 </site>
</loc>
<period>1940</period>
<people>
 <authors>
  <names>
   <name>Vincent Sheehan</name>
  </names>
  <bt>‘Personal History’</bt>
 </authors>
 <writers>
```

```
    <names>
     <kname>J.Harrison</kname>
     <name>Charles Bennett</name>
     <name>James Hilton</name>
     <name>Robert Benchley</name>
    </names>
    <pawards>
     <paw>AAN</paw>
    </pawards>
   </writers>
   <cingraphs>
    <names>
     <name>Mate</name>
    </names>
    <pawards>
     <paw>AAN</paw>
    </pawards>
   </cingraphs>
  </people>
  <notes>
   <source>
    <seen>3Jan1991, 12Jan2000</seen>
    <vt>N-HT1</vt>
   </source>
  </notes>
</film>

<film fid="H34">
 <t>Mr.~and Mrs.~Smith</t>
 <year>1941</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Eddington</pname>
  </prod>
 </prods>
 <studios>
  <studio>RKO</studio>
 </studios>
 <prcs>
  <prc>bnw</prc>
 </prcs>
 <cats>
```

```
  <cat>Romt</cat>
 </cats>
 <loc>
  <site>
   <sitename>NYC</sitename>
   <siteclass>city</siteclass>
   <siteplace>NY</siteplace>
  </site>
 </loc>
 <people>
  <visuals>
   <names>
    <kname>Polglase</kname>
   </names>
  </visuals>
  <cingraphs>
   <names>
    <name>Harry Stradling</name>
   </names>
  </cingraphs>
 </people>
 <notes>
  <source>
   <seen>23Oct1988</seen>
   <vt>N-HT4</vt>
  </source>
 </notes>
 <error>V</error>
</film>

<film fid="H35">
 <t>Suspicion</t>
 <year>1941</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Raphaelson</pname>
  </prod>
  <prod prodk="X">
   <pname>Hitchcock</pname>
  </prod>
 </prods>
 <studios>
```

```
  <studio>RKO</studio>
 </studios>
 <prcs>
  <prc>bnw</prc>
  <prc>cld</prc>
 </prcs>
 <cats>
  <cat>Susp</cat>
 </cats>
 <loc>
  <site>
   <sitedes>mansion</sitedes>
   <siteclass>estate</siteclass>
   <siteplace>GB</siteplace>
  </site>
 </loc>
 <people>
  <authors>
   <names>
    <name>Francis Iles</name>
   </names>
  </authors>
  <cingraphs>
   <names>
    <name>Harry Stradling</name>
   </names>
  </cingraphs>
 </people>
 <notes>
  <money>
   <profit>440K</profit>
  </money>
 </notes>
</film>

<film fid="H36">
 <t>Saboteur</t>
 <year>1942</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>F.Lloyd</pname>
  </prod>
```

```
  <prod prodk="X">
   <pname>Skirball</pname>
  </prod>
 </prods>
 <studios>
  <studio>Universal</studio>
 </studios>
 <prcs>
  <prc>bnw</prc>
 </prcs>
 <cats>
  <cat>Susp</cat>
 </cats>
 <awards>
  <aw>
   <awtype>H</awtype>
   <awattr>***</awattr>
  </aw>
 </awards>
 <loc>
  <site>
   <sitedes>national monuments</sitedes>
   <siteclass>am.park</siteclass>
   <siteplace>WY</siteplace>
  </site>
  <site>
   <sitename>statue of liberty</sitename>
   <siteclass>am.park</siteclass>
   <siteplace>NY</siteplace>
  </site>
 </loc>
 <people>
  <cingraphs>
   <names>
    <kname>Valentine</kname>
   </names>
  </cingraphs>
 </people>
 <notes>
  <source>
   <seen>20Sep1989</seen>
   <vt>N-HT2</vt>
  </source>
 </notes>
</film>

<film fid="H37">
```

```
<t>Shadow of a Doubt</t>
<year>1943</year>
<dirs>
 <dir dirk="R">
  <dirn>Hitchcock</dirn>
 </dir>
</dirs>
<prods>
 <prod prodk="R">
  <pname>Skirball</pname>
 </prod>
</prods>
<studios>
 <studio>Universal</studio>
</studios>
<prcs>
 <prc>bnw</prc>
</prcs>
<cats>
 <cat>Susp</cat>
</cats>
<awards>
 <aw>
  <awtype>VIP</awtype>
  <awattr>two of everything</awattr>
  <awref>[Rohmer]</awref>
 </aw>
</awards>
<loc>
 <site>
  <sitename>San Rafael</sitename>
  <siteclass>town</siteclass>
  <siteplace>CA</siteplace>
 </site>
</loc>
<period>1938</period>
<people>
 <writers>
  <names>
   <name>Thornton Wilder</name>
  </names>
 </writers>
 <cingraphs>
  <names>
   <kname>Valentine</kname>
  </names>
 </cingraphs>
```

```
  <composers>
   <names>
    <kname>Tiomkin</kname>
    <name>Previn</name>
    <name>Johann Strauss~jr.</name>
   </names>
   <bt>Merry Widow Waltz</bt>
  </composers>
 </people>
 <notes>
  <source>
   <seen>10Jul1991</seen>
  </source>
 </notes>
</film>

<film fid="H38">
 <t>Lifeboat</t>
 <year>1943</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>MacGowan</pname>
  </prod>
 </prods>
 <studios>
  <studio>Fox</studio>
 </studios>
 <prcs>
  <prc>bnw</prc>
 </prcs>
 <cats>
  <cat>Susp</cat>
 </cats>
 <awards>
  <aw>
   <awtype>H</awtype>
   <awattr>**</awattr>
  </aw>
  <aw>
   <awtype>AANdir</awtype>
  </aw>
 </awards>
```

```
 <loc>
  <site>
   <sitedes>sea</sitedes>
   <siteclass>sea</siteclass>
  </site>
 </loc>
 <people>
  <authors>
   <names>
    <kname>Steinbeck</kname>
   </names>
   <pawards>
    <paw>AAN</paw>
   </pawards>
  </authors>
  <writers>
   <names>
    <kname>Steinbeck</kname>
    <name>Jo Swerling</name>
   </names>
  </writers>
  <cingraphs>
   <names>
    <name>Glen MacWilliams</name>
   </names>
   <pawards>
    <paw>AAN</paw>
   </pawards>
  </cingraphs>
 </people>
 <notes>
  <source>
   <seen>18Sep1989</seen>
   <vt>NHT4</vt>
  </source>
 </notes>
</film>

<film fid="H39">
 <t>Spellbound</t>
 <year>1945</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
```

```
 <prod prodk="R">
  <pname>Selznick</pname>
 </prod>
</prods>
<studios>
 <studio>Selznick</studio>
 <distributor>U.A.</distributor>
</studios>
<prcs>
 <prc>bnw</prc>
</prcs>
<cats>
 <cat>Susp</cat>
</cats>
<awards>
 <aw>
  <awtype>H</awtype>
  <awattr>**</awattr>
 </aw>
 <aw>
  <awtype>AAN</awtype>
 </aw>
 <aw>
  <awtype>AANdir</awtype>
 </aw>
</awards>
<loc>
 <site>
  <sitedes>mental hospital</sitedes>
  <siteclass>hospital</siteclass>
  <siteplace>Vermont</siteplace>
 </site>
</loc>
<people>
 <authors>
  <names>
   <name>Francis Beeding</name>
  </names>
  <bt>‘‘The House of Dr.Edwardees’’</bt>
 </authors>
 <writers>
  <names>
   <kname>Hecht</kname>
  </names>
 </writers>
 <visuals>
  <names>
```

```
    <name>Salvador Dali</name>
   </names>
  </visuals>
  <cingraphs>
   <names>
    <kname>George~Barnes</kname>
   </names>
   <pawards>
    <paw>AAN</paw>
   </pawards>
  </cingraphs>
  <composers>
   <names>
    <name>Rozsa</name>
   </names>
   <pawards>
    <paw>AA</paw>
   </pawards>
  </composers>
 </people>
 <notes>
  <source>
   <seen>3Dec1989, 5May1990</seen>
  </source>
 </notes>
</film>

<film fid="H40">
 <t>Notorious</t>
 <year>1946</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Hitchcock</pname>
  </prod>
 </prods>
 <studios>
  <studio>RKO</studio>
 </studios>
 <prcs>
  <prc>bnw</prc>
 </prcs>
 <cats>
```

```
  <cat>Susp</cat>
 </cats>
 <awards>
  <aw>
   <awtype>H</awtype>
   <awattr>***</awattr>
  </aw>
  <aw>
   <awtype>AFISp</awtype>
   <awattr>38</awattr>
  </aw>
 </awards>
 <loc>
  <site>
   <sitename>Rio de Janeiro</sitename>
   <siteclass>city</siteclass>
   <siteplace>Brazil</siteplace>
  </site>
 </loc>
 <people>
  <writers>
   <names>
    <kname>Hecht</kname>
   </names>
   <pawards>
    <paw>AAN</paw>
   </pawards>
  </writers>
  <cingraphs>
   <names>
    <kname>Tetzlaff</kname>
   </names>
  </cingraphs>
  <composers>
   <names>
    <name>Roy Webb</name>
   </names>
  </composers>
 </people>
 <notes>
  <source>
   <seen>20Jan1990, 30Jun1997</seen>
  </source>
 </notes>
</film>

<film fid="H41">
```

```
<t>The Paradine Case</t>
<year>1947</year>
<dirs>
 <dir dirk="R">
  <dirn>Hitchcock</dirn>
 </dir>
</dirs>
<prods>
 <prod prodk="R">
  <pname>Selznick</pname>
 </prod>
</prods>
<studios>
 <studio>Selznick</studio>
 <distributor>U.A.</distributor>
</studios>
<prcs>
 <prc>bnw</prc>
</prcs>
<cats>
 <cat>Susp</cat>
</cats>
<awards>
 <aw>
  <awtype>H</awtype>
  <awattr>**</awattr>
 </aw>
</awards>
<loc>
 <site>
  <sitename>London</sitename>
  <siteclass>city</siteclass>
  <siteplace>England</siteplace>
 </site>
</loc>
<people>
 <authors>
  <names>
   <name>Robert Hichens</name>
  </names>
 </authors>
 <writers>
  <names>
   <name>Selznick</name>
  </names>
 </writers>
 <cingraphs>
```

```
    <names>
     <name>Lee Garmes</name>
    </names>
   </cingraphs>
   <composers>
    <names>
     <name>Waxman</name>
    </names>
   </composers>
  </people>
</film>

<film fid="H42">
 <t>Rope</t>
 <year>1948</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Bernstein</pname>
  </prod>
  <prod prodk="X">
   <pname>Hitchcock</pname>
  </prod>
 </prods>
 <studios>
  <studio>Transatlantic</studio>
  <distributor>Warners</distributor>
 </studios>
 <prcs>
  <prc>Tcol</prc>
  <prctext>first color Hitchcock</prctext>
  <prctext>all shot in long, 10 minute takes</prctext>
 </prcs>
 <cats>
  <cat>Susp</cat>
 </cats>
 <awards>
  <aw>
   <awtype>H</awtype>
   <awattr>**</awattr>
  </aw>
 </awards>
 <loc>
```

```
  <site>
   <sitedes>penthouse</sitedes>
   <siteclass>apartment</siteclass>
   <siteat>NYC</siteat>
   <siteplace>NY</siteplace>
  </site>
 </loc>
 <people>
  <cingraphs>
   <names>
    <kname>Valentine</kname>
    <name>William V. Skall</name>
   </names>
  </cingraphs>
  <composers>
   <names>
    <name>Francis Poulenc</name>
   </names>
   <bt>Mouvement Perpetuel</bt>
  </composers>
 </people>
 <notes>
  <source>
   <seen>25Nov1992</seen>
  </source>
 </notes>
</film>

<film fid="H43">
 <t>Under Capricorn</t>
 <year>1949</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Bernstein</pname>
  </prod>
  <prod prodk="X">
   <pname>Hitchcock</pname>
  </prod>
 </prods>
 <studios>
  <studio>Transatlantic</studio>
  <studio>MGM British</studio>
```

```
  <distributor>Warners</distributor>
 </studios>
 <prcs>
  <prc>Tcol</prc>
 </prcs>
 <cats>
  <cat>Dram</cat>
 </cats>
 <loc>
  <site>
   <sitename>Sidney</sitename>
   <siteclass>city</siteclass>
   <siteplace>Australia</siteplace>
  </site>
 </loc>
 <people>
  <authors>
   <names>
    <name>Helen Simpson</name>
   </names>
  </authors>
  <authors>
   <names>
    <kname>Patricia~Highsmith</kname>
   </names>
  </authors>
  <writers>
   <names>
    <name>James Bridie</name>
    <name>Hume Cronyn</name>
   </names>
  </writers>
  <cingraphs>
   <names>
    <name>Jack Cardiff</name>
    <name>Ian Craig</name>
    <name>David McNeilly</name>
   </names>
  </cingraphs>
 </people>
 <notes>
  <source>
   <seen>15May1990</seen>
  </source>
 </notes>
</film>
```

```
<film fid="H44">
 <t>Stage Fright</t>
 <alts>
  <alt>
   <altt>Die rote Lola</altt>
   <altwhy>\Ge</altwhy>
  </alt>
 </alts>
 <year>1950</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Hitchcock</pname>
  </prod>
  <prod prodk="R">
   <pname>Ahern</pname>
  </prod>
 </prods>
 <studios>
  <studio>Elstree</studio>
  <studioloc>\Ge</studioloc>
  <studioloc>\GB</studioloc>
  <distributor>Warners</distributor>
 </studios>
 <prcs />
 <cats>
  <cat>Susp</cat>
 </cats>
 <loc>
  <site>
   <siteplace>England</siteplace>
  </site>
  <site>
   <siteplace>Germany</siteplace>
  </site>
  <site>
   <siteclass>theater</siteclass>
   <siteat>London</siteat>
   <siteplace>GB</siteplace>
  </site>
 </loc>
 <people>
  <cingraphs>
```

```
    <names>
     <name>Wilkie Cooper</name>
    </names>
   </cingraphs>
  </people>
</film>

<film fid="H45">
 <t>Strangers on a Train</t>
 <year>1951</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Hitchcock</pname>
  </prod>
 </prods>
 <studios>
  <studio>Warners</studio>
 </studios>
 <prcs>
  <prc>bnw</prc>
 </prcs>
 <cats>
  <cat>Susp</cat>
 </cats>
 <awards>
  <aw>
   <awtype>H</awtype>
   <awattr>***</awattr>
  </aw>
  <aw>
   <awtype>AFISp</awtype>
   <awattr>32</awattr>
  </aw>
 </awards>
 <loc>
  <site>
   <sitename>NYC</sitename>
   <siteclass>city</siteclass>
   <siteplace>NY</siteplace>
  </site>
  <site>
   <siteclass>train</siteclass>
```

```
  </site>
  <site>
   <sitedes>merrygoround</sitedes>
   <siteclass>am.park</siteclass>
  </site>
  <site>
   <siteclass>csd</siteclass>
  </site>
 </loc>
 <people>
  <cingraphs>
   <names>
    <kname>Burks</kname>
   </names>
   <pawards>
    <paw>AAN</paw>
   </pawards>
  </cingraphs>
  <composers>
   <names>
    <kname>Tiomkin</kname>
   </names>
  </composers>
 </people>
 <notes>
  <facts>Symbols:crossings</facts>
  <source>
   <seen>10Jun1989, 2Mar1998</seen>
   <vt>NHT8</vt>
  </source>
 </notes>
</film>

<film fid="H46">
 <t>I Confess</t>
 <year>1952</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Hitchcock</pname>
  </prod>
 </prods>
 <studios>
```

```
  <studio>Warners</studio>
 </studios>
 <prcs>
  <prc>bnw</prc>
 </prcs>
 <cats>
  <cat>Susp</cat>
 </cats>
 <loc>
  <site>
   <siteclass>theater</siteclass>
   <siteat>Quebec City, Quebec</siteat>
   <siteplace>Canada</siteplace>
  </site>
 </loc>
 <people>
  <cingraphs>
   <names>
    <kname>Burks</kname>
   </names>
  </cingraphs>
 </people>
</film>

<film fid="H47">
 <t>Dial M for Murder</t>
 <year>1954</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Hitchcock</pname>
  </prod>
 </prods>
 <studios>
  <studio>Warners</studio>
 </studios>
 <prcs>
  <prc>Wcol 3D</prc>
 </prcs>
 <cats>
  <cat>Susp</cat>
 </cats>
 <awards>
```

```
  <aw>
   <awtype>H</awtype>
   <awattr>**</awattr>
  </aw>
 </awards>
 <loc>
  <site>
   <sitename>London</sitename>
   <siteclass>city</siteclass>
   <siteplace>GB</siteplace>
  </site>
 </loc>
 <people>
  <cingraphs>
   <names>
    <kname>Burks</kname>
   </names>
  </cingraphs>
  <composers>
   <names>
    <kname>Tiomkin</kname>
   </names>
  </composers>
 </people>
 <notes>
  <source>
   <seen>3Dec1989</seen>
  </source>
 </notes>
</film>

<film fid="H48">
 <t>Rear Window</t>
 <year>1954</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Hitchcock</pname>
  </prod>
 </prods>
 <studios>
  <studio>Paramount</studio>
 </studios>
```

```
<prcs>
 <prc>Tcol</prc>
</prcs>
<cats>
 <cat>Susp</cat>
</cats>
<awards>
 <aw>
  <awtype>H</awtype>
  <awattr>***</awattr>
 </aw>
 <aw>
  <awtype>AANdir</awtype>
 </aw>
 <aw>
  <awtype>AFISp</awtype>
  <awattr>14</awattr>
 </aw>
</awards>
<loc>
 <site>
  <siteclass>town</siteclass>
  <siteat>East</siteat>
  <siteplace>USA</siteplace>
 </site>
</loc>
<people>
 <authors>
  <names>
   <kname>Woolrich</kname>
  </names>
 </authors>
 <writers>
  <names>
   <kname>J.M.Hayes</kname>
  </names>
  <pawards>
   <paw>AAN</paw>
  </pawards>
 </writers>
 <cingraphs>
  <names>
   <kname>Burks</kname>
  </names>
  <pawards>
   <paw>AAN</paw>
  </pawards>
```

```
  </cingraphs>
 </people>
 <notes>
  <source>
   <vt>NHC3; HT8, inc</vt>
  </source>
 </notes>
</film>

<film fid="H49">
 <t>To Catch a Thief</t>
 <year>1955</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Hitchcock</pname>
  </prod>
 </prods>
 <studios>
  <studio>Paramount</studio>
 </studios>
 <prcs>
  <prc>Tcol</prc>
 </prcs>
 <cats>
  <cat>Susp</cat>
 </cats>
 <loc>
  <site>
   <sitename>Riviera</sitename>
   <siteclass>resort</siteclass>
   <siteplace>France</siteplace>
  </site>
 </loc>
 <people>
  <authors>
   <names>
    <name>David Dodge</name>
   </names>
  </authors>
  <writers>
   <names>
    <kname>J.M.Hayes</kname>
```

```
    </names>
   </writers>
   <cingraphs>
    <names>
     <kname>Burks</kname>
    </names>
   </cingraphs>
  </people>
</film>

<film fid="H50">
 <t>The Trouble with Harry</t>
 <year>1956</year>
 <dirs>
  <dir dirk="R">
   <dirn>Hitchcock</dirn>
  </dir>
 </dirs>
 <prods>
  <prod prodk="R">
   <pname>Hitchcock</pname>
  </prod>
 </prods>
 <studios>
  <studio>Paramount</studio>
 </studios>
 <prcs>
  <prc>Tcol</prc>
 </prcs>
 <cats>
  <cat>Susp</cat>
 </cats>
 <loc>
  <site>
   <siteclass>csd</siteclass>
   <siteplace>Vt</siteplace>
  </site>
 </loc>
 <people>
  <writers>
   <names>
    <kname>J.M.Hayes</kname>
   </names>
  </writers>
  <cingraphs>
   <names>
    <kname>Burks</kname>
```

```
    </names>
   </cingraphs>
  </people>
  <notes>
   <source>
    <seen>1957, 1989</seen>
   </source>
  </notes>
 </film>
</movies>
```

Figura A.1: Documento sin minimizar (registro50.sgml).

A.2. Documento minimizado

```
<!DOCTYPE movies SYSTEM "dtd_modificada.dtd">
<film fid=H1>Always Tell Your Wife
 <year>1922
 <dir R>Se.Hicks
 <dir R>Hitchcock
 <prods<prod R>Lasky
 <studios>Famous
 <prcs>sbw
 <cats>Dram
<film fid=H2>Number Thirteen
 <year>1922
 <dir R>Hitchcock
 <prods<prod R>Hitchcock
 <studios>Islington
 <distributor>Famous
 <prcs>sbw
 <prctext>unfinished
<film fid=H3>Woman to Woman
 <year>1922
 <dir R>Hitchcock
 <dir>Cutts
 <prods<prod R>Balcon
 <studios>B-S-F
 <distributor>Wardour
 <prcs>sbw
 <cats>Dram
 <loc<siteplace>GB
 <error>same(GCt27), y(1926)
<film fid=H4>The Passionate Adventure
 <year>1924
```

```
 <dir R>Hitchcock
 <dir>Cutts
 <prods<prod R>Balcon
 <studios>Gainsborough
 <distributor>GaumontD
 <prcs>sbw
<film fid=H5>The Blackguard
 <year>1925
 <dir R>Hitchcock
 <dir>Cutts
 <prods<prod R>Balcon
 <studios>UFA
 <distributor>Wardour
 <prcs>sbw
<film fid=H6>The Pleasure Garden
 <year>1925
 <dir R>Hitchcock
 <prods<prod R>Balcon
 <studios>Gainsborough and Emelka
 <distributor>Wardour
 <prcs>sbw
 <cats>Dram
 <awards>H
 <awattr>0
 <loc<sitename>Pleasure Garden
 <siteclass>theater
 <people<authors<name>Oliver Sandys
 <writers<name>Eliot Stannard
 <cingraphs<kname>Ventimiglia
<film fid=H7>The Mountain Eagle
 <alts>Fear O'God
 <altwhy>USA
 <year>1926
 <dir R>Hitchcock
 <prods<prod R>Balcon
 <studios>Gainsborough
 <>Emelka
 <distributor>Wardour
 <prcs>sbw
 <prctext>no prints left
 <cats>Dram
 <awards>H
 <awattr>0
 <people<writers<name>Eliot Stannard
 <cingraphs<kname>Ventimiglia
<film fid=H8>The Lodger: A Story of The London Fog
 <alts>The Case of Jonathan Drew
```

```
 <altwhy>USA
 <year>1926
 <dir R>Hitchcock
 <prods<prod R>Balcon
 <studios>Gainsborough
 <distributor>Wardour
 <prcs>sbw
 <cats>Susp
 <loc<sitename>London
 <siteclass>city
 <siteplace>GB
 <people<authors<name>Mary Belloc Lowndes
 <bt>The Lodger
 <writers<name>Hitchcock
 <name>Eliot Stannard
 <cingraphs<kname>Ventimiglia
 <notes<money<cost>12K
 <moneynotes>in UK lbs
 <facts>Topic:Jack the Ripper
<film fid=H9>Downhill
 <alts>When Boys Leave Home
 <year>1927
 <dir R>Hitchcock
 <prods<prod R>Balcon
 <studios>Islington
 <distributor>Wardour
 <prcs>sbw
 <cats>Susp
 <people<authors<name>Constance Collier
 <writers<name>Eliot Stannard
 <cingraphs<kname>Ventimiglia
 <error>W(Ivor Novello)
<film fid=H10>Easy Virtue
 <year>1927
 <dir R>Hitchcock
 <prods<prod R>Balcon
 <studios>Gainsborough
 <distributor>Wardour
 <prcs>sbw
 <cats>Susp
 <loc<sitedes>artist's studio
 <siteclass>work
 <siteplace>England
 <site<siteclass>hotel
 <siteat>Riviera
 <siteplace>France
 <site<sitedes>country mansion
```

```
 <siteclass>estate
 <siteplace>England
 <people<authors<name>Noel Coward
 <writers<name>Eliot Stannard
 <cingraphs<name>Claude McDonnell
 <notes<source<seen>11Feb2000
<film fid=H11>The Ring
 <year>1927
 <dir R>Hitchcock
 <prods<prod R>J.Maxwell
 <studios>BIP Elstree
 <distributor>Wardour
 <prcs>sbw
 <cats>Susp
 <loc<sitedes>boxing rings
 <siteclass>sports
 <siteplace>England
 <site<sitename>Albert Hall
 <siteclass>theater
 <siteat>London
 <siteplace>GB
 <people<writers<name>Hitchcock
 <name>Alma Reveille
 <cingraphs<name>Claude McDonnell
 <notes<source<seen>11Feb2000
<film fid=H12>The Farmer's Wife
 <year>1928
 <dir R>Hitchcock
 <prods<prod R>J.Maxwell
 <studios>BIP Elstree
 <distributor>Wardour
 <prcs>sbw
 <cats>Susp
 <people<authors<name>Eden Philpotts
 <writers<name>Hitchcock
 <cingraphs<name>J.Cox
<film fid=H13>Champagne
 <year>1928
 <dir R>Hitchcock
 <prods<prod R>J.Maxwell
 <studios>BIP Elstree
 <distributor>Wardour
 <prcs>sbw
 <cats>Romt
 <people<authors<name>Walter Mycroft
 <writers<name>Eliot Stannard
 <cingraphs<name>J.Cox
```

```
<film fid=H14>Harmony Heaven
 <year>1929
 <dir R>Hitchcock
 <dir>Thomas Bentley
 <studios>BIP
 <distributor>France-Societ\'e des Cin\'e-Romans
 <prcs>col
 <cats>Susp
 <awards>H
 <awattr>0
 <people<writers<name>Arthur Wimperis
 <name>Randall Faye
 <cingraphs<name>J.Cox
 <composers<name>Eddie Pola
 <name>Edward Brandt
<film fid=H15>The Manxman
 <year>1929
 <dir R>Hitchcock
 <prods<prod R>J.Maxwell
 <studios>BIP Elstree
 <distributor>Wardour
 <prcs>sbw
 <cats>Susp
 <awards>H
 <awattr>0
 <loc<siteclass>csd
 <siteat>Scotland
 <siteplace>GB
 <people<authors<name>Hall Caine
 <writers<name>Eliot Stannard
 <cingraphs<name>J.Cox
<film fid=H16>Blackmail
 <year>1929
 <dir R>Hitchcock
 <prods<prod R>J.Maxwell
 <studios>BIP Elstree
 <distributor>Wardour
 <prcs>bnw
 <prctext>first sound, added
 <cats>Susp
 <loc<sitename>British Museum
 <siteclass>museum
 <siteat>London
 <siteplace>GB
 <people<writers<name>Hitchcock
 <name>Benn W. Levy
 <name>Charles Bennett
```

```
 <cingraphs<name>J.Cox
 <notes<source<seen>9Apr1990
<film fid=H17>Elstree Calling
 <year>1929
 <dir R>Brunel
 <dir>Hitchcock
 <dir>Andr\'e Charlot
 <dir>Jack Hulbert
 <dir>Paul Murray
 <studios>BIP Elstree
 <prcs>bnw
 <cats>Musc
 <loc<sitedes>studio
 <siteclass>work
 <people<writers<name>Ivor Novello
 <notes<facts>Brunel was the supervisor
<film fid=H18>Juno And The Paycock
 <year>1930
 <dir R>Hitchcock
 <prods<prod R>J.Maxwell
 <studios>BIP Elstree
 <distributor>Wardour
 <prcs>bnw
 <cats>Susp
 <awards>H
 <awattr>*
 <loc<sitename>Dublin
 <siteclass>city
 <siteplace>Ireland
 <people<authors<name>Sean O'Casey
 <writers<name>Hitchcock
 <name>Alma Reville
 <cingraphs<name>J.Cox
<film fid=H19>Murder!
 <year>1930
 <dir R>Hitchcock
 <prods<prod R>J.Maxwell
 <prod X>Alfred Abel
 <prodnote>Abel produced the German version
 <studios>BIP Elstree
 <distributor>Wardour
 <prcs>bnw
 <cats>Myst
 <awards>H
 <awattr>**
 <loc<siteclass>theater
 <siteat>London
```

```
 <siteplace>GB
 <people<authors
 <name>Clarence Dane
 <name>Helen Simpson
 <bt>Enter Sir John
 <writers<kname>Hitchcock
 <name>Alma Reville
 <visuals<name>Winifred Ashton
 <cingraphs<name>J.Cox
 <notes<source<seen>10apr1990
 <error>V
<film fid=H20>The Skin Game
 <year>1931
 <dir R>Hitchcock
 <prods<prod R>J.Maxwell
 <studios>BIP Elstree
 <distributor>Wardour
 <prcs>bnw
 <cats>Dram
 <people<authors<name>John Galsworthy
 <writers<name>Hitchcock
 <name>Alma Reville
 <cingraphs<name>J.Cox
 <name>Charles Martin
 <error>C
<film fid=H21>Rich and Strange
 <alts>East of Shanghai
 <altwhy>USA
 <year>1932
 <dir R>Hitchcock
 <prods<prod R>J.Maxwell
 <studios>BIP Elstree
 <distributor>Wardour
 <prcs>bnw
 <cats>susp
 <loc<sitename>London
 <siteclass>city
 <siteplace>England
 <site<sitename>Paris
 <siteclass>city
 <siteplace>France
 <site<sitename>Marseille
 <siteclass>city
 <siteplace>France
 <site<siteplace>Singapore
 <site<siteclass>ship
 <siteat>sea
```

```
 <siteplace>Pacific
 <people<authors<name>Dale Collins
 <writers<name>Alma Reville
 <name>Val Valentine
 <cingraphs<name>J.Cox
 <name>Charles Martin
 <notes<source<seen>16apr1990
 <error>C
<film fid=H22>Number Seventeen
 <year>1932
 <dir R>Hitchcock
 <prods<prod R>J.Maxwell
 <studios>BIP Elstree
 <distributor>Wardour
 <prcs>bnw
 <cats>Myst
 <awards>H
 <awattr>*
 <loc<sitename>London
 <siteclass>city
 <siteplace>England
 <site<siteclass>train
 <siteplace>England
 <site<sitename>Dover
 <siteclass>town
 <siteplace>England
 <people<authors<name>Jefferson Farejon
 <writers<name>Jefferson Farejon
 <name>Rodney Auckland
 <cingraphs<name>J.Cox
 <notes<source<seen>16apr1990
 <error>W
<film fid=H23>Lord Camber's Ladies
 <year>1932
 <dir R>Hitchcok
 <dir>Benn~Levy
 <prods<prod R>Hitchcock
 <studios>BIP Elstree
 <distributor>Wardour
 <prcs>bnw
 <cats>Susp
 <loc<sitedes>castle
 <siteclass>palace
 <siteplace>GB
 <people<authors<name>H.A. Vachell
 <bt>The Case of Lady Camber
 <writers<name>Hitchcock
```

```
 <name>Jefferson Farejon
<film fid=H24>Waltzes From Vienna
 <alts>Strauss's Great Waltz
 <altwhy>USA
 <year>1933
 <dir R>Hitchcock
 <prods<prod N>Tom Arnold
 <studios>Lime Grove
 <distributor>G.F.D.
 <prcs>bnw
 <cats>Romt
 <>Comd
 <awards>H
 <awattr>0
 <aw>VIP
 <awattr>least favorite of Hitchcock
 <loc<sitename>Vienna
 <siteclass>city
 <siteplace>Austria
 <people<writers<name>Alma Reveille
 <name>Bolton
 <composers<name>Johann Strauss~sr.
 <name>Johann Strauss~jr.
<film fid=H25>The Man Who Knew Too Much
 <year>1934
 <dir R>Hitchcock
 <prods<prod R>Balcon
 <prod X>Montagu
 <studios>Gaumont Lime Grove
 <distributor>G.F.D.
 <prcs>bnw
 <cats>Susp
 <awards>H
 <awattr>***
 <loc<sitename>St.Moritz
 <siteclass>mountains
 <siteplace>Switzerland
 <site<sitename>London
 <siteclass>city
 <siteplace>GB
 <people<authors<name>D.B. Wyndham-Lewis
 <name>Edwin Greenwood
 <writers<name>A.R.Rawlinson
 <name>Charles Bennett
 <name>D.B. Wyndham-Lewis
 <name>Edwin Greenwood
 <cingraphs<name>Curt Courant
```

```
 <composers<kname>Louis~Levy
 <name>Arthur Benjamin
<film fid=H26>The 39 Steps
 <year>1935
 <dir R>Hitchcock
 <prods<prod R>Balcon
 <prod X>Montagu
 <studios>Gaumont Lime Grove
 <distributor>G.F.D.
 <prcs>bnw
 <cats>Susp
 <awards>H
 <awattr>****
 <loc<siteclass>theater
 <siteat>London
 <siteplace>GB
 <people<authors
 <name>John Buchan
 <writers<name>Charles Bennett
 <name>Alma Reveille
 <cingraphs<name>Bernard Knowles
 <composers<name>Louis Levy
<film fid=H27>Secret Agent
 <year>1936
 <dir R>Hitchcock
 <prods<prod R>Balcon
 <prod X>Montagu
 <studios>Gaumont Lime Grove
 <distributor>G.F.D.
 <prcs>bnw
 <cats>Susp
 <loc<sitename>London
 <siteclass>city
 <siteplace>England
 <site<sitedes>chocolate factory
 <siteclass>work
 <siteplace>Switzerland
 <people<authors
 <name>S.Maugham
 <bt>Ashenden
 <writers<name>Charles Bennett
 <name>Campbell Dixon
 <cingraphs<name>Bernard Knowles
 <error>W
<film fid=H28>Sabotage
 <alts>The Woman Alone
 <altwhy>USA
```

```
 <year>1936
 <dir R>Hitchcock
 <prods<prod R>Balcon
 <prod X>Montague
 <studios>Lime Grove
 <distributor>G.F.D.
 <prcs>bnw
 <cats>Susp
 <awards>H
 <awattr>***
 <loc<siteclass>theater
 <siteat>London
 <siteplace>GB
 <people<authors<name>Conrad
 <bt>The Secret Agent
 <writers<name>Charles Bennett
 <cingraphs<name>Bernard Knowles
<film fid=H29>Young and Innocent
 <alts>The Girl Was Young
 <altwhy>USA
 <year>1938
 <dir R>Hitchcock
 <prods<prod R>Black
 <studios>Lime Grove and Pinewood
 <distributor>G.F.D.
 <prcs>bnw
 <cats>Susp
 <people<authors<name>Josephine Tey
 <writers<name>Charles Bennett
 <name>Alma Reveille
 <cingraphs<name>Bernard Knowles
<film fid=H30>The Lady Vanishes
 <year>1938
 <dir R>Hitchcock
 <prods<prod R>Black
 <studios>Lime Grove Gainsborough
 <distributor>MGM
 <prcs>bnw
 <cats>Susp
 <loc<siteclass>train
 <site<siteplace>Austria
 <people<authors<name>Ethel Lina White
 <bt>‘The Wheel Spins’
 <writers<name>Gilliat
 <name>Launder
 <cingraphs<name>J.Cox
 <notes<source<seen>11Jul1988
```

```
<film fid=H31>Jamaica Inn
 <year>1939
 <dir R>Hitchcock
 <prods<prod R>Pommer
 <prod X>Laughton
 <studios>Elstree
 <distributor>Associated British
 <distributor>Paramount
 <prcs>bnw
 <cats>Dram
 <awards>W50
 <loc<sitedes>seashore
 <siteclass>resort
 <siteat>Cornwall
 <siteplace>GB
 <people<cingraphs<name>Bernard Knowles
 <name>Harry Stradling
 <error>C
<film fid=H32>Rebecca
 <year>1940
 <dir R>Hitchcock
 <prods<prod R>Selznick
 <studios>Selznick
 <distributor>U.A.
 <prcs>bnw
 <cats>Dram
 <awards>AA
 <aw>AANdir
 <aw>H
 <awattr>****
 <loc<sitename>Mandalay
 <sitedes>castle
 <siteclass>estate
 <siteplace>GB
 <people<authors
 <name>Daphne duMaurier
 <writers<kname>J.Harrison
 <name>Robert E. Sherwood
 <pawards>AAN
 <visuals<kname>Lyle~Wheeler
 <cingraphs
 <kname>George~Barnes
 <pawards>AA
 <composers<name>Waxman
 <pawards>AAN
 <notes<source<seen>1989, 25Jan2000
 <vt>N-HT5
```

```
<film fid=H33>Foreign Correspondent
 <year>1940
 <dir R>Hitchcock
 <prods<prod R>Wanger
 <studios>U.A.
 <prcs>bnw
 <cats>Susp
 <awards>H
 <awattr>****
 <aw>AAN
 <loc<sitename>London
 <siteclass>city
 <siteplace>England
 <site<siteplace>Netherlands
 <site<sitedes>flying boat
 <siteclass>airplane
 <siteplace>Atlantic
 <period>1940
 <people<authors<name>Vincent Sheehan
 <bt>‘Personal History’
 <writers<kname>J.Harrison
 <name>Charles Bennett
 <name>James Hilton
 <name>Robert Benchley
 <pawards>AAN
 <cingraphs<name>Mate
 <pawards>AAN
 <notes<source<seen>3Jan1991, 12Jan2000
 <vt>N-HT1
<film fid=H34>Mr.~and Mrs.~Smith
 <year>1941
 <dir R>Hitchcock
 <prods<prod R>Eddington
 <studios>RKO
 <prcs>bnw
 <cats>Romt
 <loc<sitename>NYC
 <siteclass>city
 <siteplace>NY
 <people<visuals<kname>Polglase
 <cingraphs<name>Harry Stradling
 <notes<source<seen>23Oct1988
 <vt>N-HT4
 <error>V
<film fid=H35>Suspicion
 <year>1941
 <dir R>Hitchcock
```

```
 <prods<prod R>Raphaelson
 <prod X>Hitchcock
 <studios>RKO
 <prcs>bnw
 <>cld
 <cats>Susp
 <loc<sitedes>mansion
 <siteclass>estate
 <siteplace>GB
 <people<authors
 <name>Francis Iles
 <cingraphs<name>Harry Stradling
 <notes<money<profit>440K
<film fid=H36>Saboteur
 <year>1942
 <dir R>Hitchcock
 <prods<prod R>F.Lloyd
 <prod X>Skirball
 <studios>Universal
 <prcs>bnw
 <cats>Susp
 <awards>H
 <awattr>***
 <loc<sitedes>national monuments
 <siteclass>am.park
 <siteplace>WY
 <site<sitename>statue of liberty
 <siteclass>am.park
 <siteplace>NY
 <people<cingraphs<kname>Valentine
 <notes<source<seen>20Sep1989
 <vt>N-HT2
<film fid=H37>Shadow of a Doubt
 <year>1943
 <dir R>Hitchcock
 <prods<prod R>Skirball
 <studios>Universal
 <prcs>bnw
 <cats>Susp
 <awards>VIP
 <awattr>two of everything
 <awref>[Rohmer]
 <loc<sitename>San Rafael
 <siteclass>town
 <siteplace>CA
 <period>1938
 <people<writers<name>Thornton Wilder
```

```
 <cingraphs<kname>Valentine
 <composers<kname>Tiomkin
 <name>Previn
 <name>Johann Strauss~jr.
 <bt>Merry Widow Waltz
 <notes<source<seen>10Jul1991
<film fid=H38>Lifeboat
 <year>1943
 <dir R>Hitchcock
 <prods<prod R>MacGowan
 <studios>Fox
 <prcs>bnw
 <cats>Susp
 <awards>H
 <awattr>**
 <aw>AANdir
 <loc<sitedes>sea
 <siteclass>sea
 <people<authors<kname>Steinbeck
 <pawards>AAN
 <writers<kname>Steinbeck
 <name>Jo Swerling
 <cingraphs<name>Glen MacWilliams
 <pawards>AAN
 <notes<source<seen>18Sep1989
 <vt>NHT4
<film fid=H39>Spellbound
 <year>1945
 <dir R>Hitchcock
 <prods<prod R>Selznick
 <studios>Selznick
 <distributor>U.A.
 <prcs>bnw
 <cats>Susp
 <awards>H
 <awattr>**
 <aw>AAN
 <aw>AANdir
 <loc<sitedes>mental hospital
 <siteclass>hospital
 <siteplace>Vermont
 <people<authors
 <name>Francis Beeding
 <bt>‘‘The House of Dr.Edwardees’’
 <writers<kname>Hecht
 <visuals<name>Salvador Dali
 <cingraphs<kname>George~Barnes
```

```
 <pawards>AAN
 <composers<name>Rozsa
 <pawards>AA
 <notes<source<seen>3Dec1989, 5May1990
<film fid=H40>Notorious
 <year>1946
 <dir R>Hitchcock
 <prods<prod R>Hitchcock
 <studios>RKO
 <prcs>bnw
 <cats>Susp
 <awards>H
 <awattr>***
 <aw>AFISp
 <awattr>38
 <loc<sitename>Rio de Janeiro
 <siteclass>city
 <siteplace>Brazil
 <people<writers<kname>Hecht
 <pawards>AAN
 <cingraphs<kname>Tetzlaff
 <composers<name>Roy Webb
 <notes<source<seen>20Jan1990, 30Jun1997
<film fid=H41>The Paradine Case
 <year>1947
 <dir R>Hitchcock
 <prods<prod R>Selznick
 <studios>Selznick
 <distributor>U.A.
 <prcs>bnw
 <cats>Susp
 <awards>H
 <awattr>**
 <loc<sitename>London
 <siteclass>city
 <siteplace>England
 <people<authors<name>Robert Hichens
 <writers<name>Selznick
 <cingraphs<name>Lee Garmes
 <composers<name>Waxman
<film fid=H42>Rope
 <year>1948
 <dir R>Hitchcock
 <prods<prod R>Bernstein
 <prod X>Hitchcock
 <studios>Transatlantic
 <distributor>Warners
```

```
 <prcs>Tcol
 <prctext>first color Hitchcock
 <prctext>all shot in long, 10 minute takes
 <cats>Susp
 <awards>H
 <awattr>**
 <loc<sitedes>penthouse
 <siteclass>apartment
 <siteat>NYC
 <siteplace>NY
 <people<cingraphs
 <kname>Valentine
 <name>William V. Skall
 <composers<name>Francis Poulenc
 <bt>Mouvement Perpetuel
 <notes<source<seen>25Nov1992
<film fid=H43>Under Capricorn
 <year>1949
 <dir R>Hitchcock
 <prods<prod R>Bernstein
 <prod X>Hitchcock
 <studios>Transatlantic
 <>MGM British
 <distributor>Warners
 <prcs>Tcol
 <cats>Dram
 <loc<sitename>Sidney
 <siteclass>city
 <siteplace>Australia
 <people<authors<name>Helen Simpson
 <authors<kname>Patricia~Highsmith
 <writers<name>James Bridie
 <name>Hume Cronyn
 <cingraphs<name>Jack Cardiff
 <name>Ian Craig
 <name>David McNeilly
 <notes<source<seen>15May1990
<film fid=H44>Stage Fright
 <alts>Die rote Lola
 <altwhy>\Ge
 <year>1950
 <dir R>Hitchcock
 <prods<prod R>Hitchcock
 <prod R>Ahern
 <studios>Elstree
 <studioloc>\Ge
 <studioloc>\GB
```

```
 <distributor>Warners
 <cats>Susp
 <loc<siteplace>England
 <site<siteplace>Germany
 <site<siteclass>theater
 <siteat>London
 <siteplace>GB
 <people<cingraphs
 <name>Wilkie Cooper
<film fid=H45>Strangers on a Train
 <year>1951
 <dir R>Hitchcock
 <prods<prod R>Hitchcock
 <studios>Warners
 <prcs>bnw
 <cats>Susp
 <awards>H
 <awattr>***
 <aw>AFISp
 <awattr>32
 <loc<sitename>NYC
 <siteclass>city
 <siteplace>NY
 <site<siteclass>train
 <site<sitedes>merrygoround
 <siteclass>am.park
 <site<siteclass>csd
 <people<cingraphs<kname>Burks
 <pawards>AAN
 <composers<kname>Tiomkin
 <notes<facts>Symbols:crossings
 <source<seen>10Jun1989, 2Mar1998
 <vt>NHT8
<film fid=H46>I Confess
 <year>1952
 <dir R>Hitchcock
 <prods<prod R>Hitchcock
 <studios>Warners
 <prcs>bnw
 <cats>Susp
 <loc<siteclass>theater
 <siteat>Quebec City, Quebec
 <siteplace>Canada
 <people<cingraphs<kname>Burks
<film fid=H47>Dial M for Murder
 <year>1954
 <dir R>Hitchcock
```

```
 <prods<prod R>Hitchcock
 <studios>Warners
 <prcs>Wcol 3D
 <cats>Susp
 <awards>H
 <awattr>**
 <loc<sitename>London
 <siteclass>city
 <siteplace>GB
 <people<cingraphs<kname>Burks
 <composers<kname>Tiomkin
 <notes<source<seen>3Dec1989
<film fid=H48>Rear Window
 <year>1954
 <dir R>Hitchcock
 <prods<prod R>Hitchcock
 <studios>Paramount
 <prcs>Tcol
 <cats>Susp
 <awards>H
 <awattr>***
 <aw>AANdir
 <aw>AFISp
 <awattr>14
 <loc<siteclass>town
 <siteat>East
 <siteplace>USA
 <people<authors<kname>Woolrich
 <writers<kname>J.M.Hayes
 <pawards>AAN
 <cingraphs<kname>Burks
 <pawards>AAN
 <notes<source<vt>NHC3; HT8, inc
<film fid=H49>To Catch a Thief
 <year>1955
 <dir R>Hitchcock
 <prods<prod R>Hitchcock
 <studios>Paramount
 <prcs>Tcol
 <cats>Susp
 <loc<sitename>Riviera
 <siteclass>resort
 <siteplace>France
 <people<authors<name>David Dodge
 <writers<kname>J.M.Hayes
 <cingraphs<kname>Burks
<film fid=H50>The Trouble with Harry
```

```
<year>1956
<dir R>Hitchcock
<prods<prod R>Hitchcock
<studios>Paramount
<prcs>Tcol
<cats>Susp
<loc<siteclass>csd
<siteplace>Vt
<people<writers<kname>J.M.Hayes
<cingraphs<kname>Burks
<notes<source<seen>1957, 1989
```

Figura A.2: Contenido del archivo registro50_min_syo.sgml (uso de todas las minimizaciones permitidas por Shorttag y Omittag en cincuenta registros).

Apéndice B

Programa desarrollado para realizar las medidas

B.1. Clase DatosFichero

```
import java.io.FileReader;
import java.io.IOException;
class DatosFichero{
 private String ruta;
  private boolean minimizado;

  DatosFichero(String ruta, boolean minimizado){
    this.ruta=ruta;
    this.minimizado=minimizado;
  }

  public long cuentaCaracteres()throws IOException{
    long caracteres=0;
    int c=0;
    FileReader f=new FileReader(ruta);

    while((c=f.read())!=-1){
      if((char)c!='\n'&& (char)c!='\r' && (char)c!='\t' && (char)c!=' ')
        // No tenemos en cuenta caracteres de nueva línea,
        // retornos de carro, tabuladores ni espacios en blanco.
        caracteres++;
    }
    f.close();
    return caracteres;
  }

  // El siguiente método cuenta los caracteres del archivo que corresponden a
  // etiquetas
  public long cuentaCaracteresEtiquetas()throws IOException{
```

```
        boolean etiqueta=false;
        long caracteres_etiquetas=0;
        long caracteres_etiqueta_actual=0;
        FileReader f=new FileReader(ruta);
        int c;

        if(!minimizado){
          while((c=f.read())!=-1){
            switch(c){
            case '<':
              if(!etiqueta){
                etiqueta=true;
                caracteres_etiqueta_actual++;
              }
              break;
          case '>':
              if(etiqueta){
                etiqueta=false;
                caracteres_etiqueta_actual++;
                caracteres_etiquetas+=caracteres_etiqueta_actual;
                caracteres_etiqueta_actual=0;
              }
              break;
            default:
              if(etiqueta && (char)c!='\n'&& (char)c!='\r' && (char)c!='\t'
                && (char)c!=' ')
                caracteres_etiqueta_actual++;
            break;
            }
          }
          f.close();
          return caracteres_etiquetas;
        }else
          // no se cuentan los caracteres correspondientes a etiquetas de
          // un archivo minimizado, se calculará restando.
          f.close();


    return -1;


      }
     }
```

Figura B.1: Archivo DatosFichero.java.

B.2. Clase ComparaCaracteres

```
import java.io.File;
import java.io.IOException;
import java.text.DecimalFormat;
import java.util.ArrayList;
class ComparaCaracteres {

  public static void main(String[] args)throws IOException {

    long caracteresCompleto;
    long caracteresMinimizado;
    long caracteresEtiquetasCompleto;
    long caracteresEtiquetasMinimizado;
    double reduccionArchivo;
    double reduccionEtiquetas;

    ArrayList ficheros_completos=new ArrayList();
    ArrayList ficheros_minimizados=new ArrayList();

    String rutaFicheroCompleto=null;
    String rutaficherominimizado=null;

    DecimalFormat df = new DecimalFormat("0.00");

   if(args.length!=1){
     System.err.println("argumentos incorrectos");
    }else{
     File f=new File(args[0]);

     // Se comprueba que el argumento es un directorio
     if(f.isDirectory()){

       // Se obtienen los archivos contenidos en el directorio.
       String lista_ficheros[]=f.list();

          // Se examina cada fichero
       for(int i=0;i<(lista_ficheros.length);i++){

         File f1=new File(args[0]+"/"+lista_ficheros[i]);

         // Se comprueba que es un fichero sgml y que no es otro directorio.
         if(!f1.isDirectory()&& lista_ficheros[i].indexOf(".sgml")!=-1){

           // Si es un archivo válido se almacena en una lista de ficheros
           // completos o minimizados según el caso.
           if(lista_ficheros[i].indexOf("_min_")!=-1){
```

```
      // Es un archivo que ha sido minimizado.
      ficheros_minimizados.add(lista_ficheros[i]);

    }else{
      // Es un archivo que no ha sido minimizado.
      ficheros_completos.add(lista_ficheros[i]);
 }
  }else
    System.out.println(lista_ficheros[i] + " es un directorio o " +
        "un archivo no válido");
}
for(int w=0;w<ficheros_completos.size();w++){
  rutaFicheroCompleto=args[0]+"\\"+ficheros_completos.get(w);
  DatosFichero datosfichero1
  =new DatosFichero(rutaFicheroCompleto,false);

  caracteresCompleto=datosfichero1.cuentaCaracteres();
  caracteresEtiquetasCompleto
  =datosfichero1.cuentaCaracteresEtiquetas();
  System.out.println("\n"+"Archivo completo: "+rutaFicheroCompleto);

  System.out.println("El archivo tiene "+caracteresEtiquetasCompleto
      +" caracteres de etiquetas y "+caracteresCompleto
      +" caracteres en total.\n");

System.out.println("Archivos obtenidos al minimizar "
      +ficheros_completos.get(w)+":");

  String aux=(String)ficheros_completos.get(w);
  String []campos_completo=aux.split("\\.");

  int z=0;
  for(int v=0;v<ficheros_minimizados.size();v++){

    String aux2=(String)ficheros_minimizados.get(v);
    String []campos_minimizado=aux2.split("_");
    if(campos_minimizado[0].equals(campos_completo[0])){
      // Se preparan las rutas de los minimizados
      // que correspondan a cada fichero completo.
      z++;
      rutaficherominimizado
      =args[0]+"\\" +ficheros_minimizados.get(v);

      DatosFichero datosfichero2
      =new DatosFichero(rutaficherominimizado,true);

 try {
```

```
                caracteresMinimizado=datosfichero2.cuentaCaracteres();

                // Los caracteres en que se ha reducido tienen que ser
                // caracteres de etiquetas, restando a las etiquetas del
                // completo las etiquetas en que se han reducido debe
                // obtenerse los caracteres de etiquetas que le quedan al
                // minimizado.

                caracteresEtiquetasMinimizado=caracteresEtiquetasCompleto

               -(caracteresCompleto-caracteresMinimizado);

                reduccionArchivo=100.0-((caracteresMinimizado*100.0)
                    /caracteresCompleto);
                reduccionEtiquetas=100.0-((caracteresEtiquetasMinimizado*100.0)
                    /caracteresEtiquetasCompleto);

                System.out.println("\n"+"Archivo Minimizado "+z+": "
                    + rutaficherominimizado);
            System.out.println("El archivo tiene "
                    + caracteresEtiquetasMinimizado
                    +" caracteres de etiquetas y "
                    + caracteresMinimizado +" caracteres en total.");

                System.out.println("Los caracteres de etiquetas "
                    +"se han reducido un "+ df.format(reduccionEtiquetas)
                    + " % y los caracteres totales un "
                    + df.format(reduccionArchivo)+" %.");

               } catch (IOException e) {
               System.out.println("Error durante el acceso a los archivos.");
               }

              }
             }if(z==0)
              System.out.println("\n"+"No se encuentran archivos minimizados" +
                  " para este archivo completo.");
          }
         }else{
          System.out.println("El argumento especificado no es un directorio");
         }
        }
       }
      }
```

Figura B.2: Archivo ComparaCaracteres.java.

Apéndice C

Glosario de términos y caracteres delimitadores

C.1. Glosario de términos usados

Término	Significado	Descripción
DTD	Document type definition	Definición de tipo de documento.
ESIS	Element Structure Information Set	Información del documento con la estructura que se haya definido
GI	Generic identifier	Identificador genérico.
GML	Generalized Markup Language	Lenguaje de marcado generalizado, precursor de SGML.
HTML	HyperText Markup Language	Lenguaje de Marcas de Hipertexto
JSON	JavaScript Object Notation	Notación de objetos de JavaScript
LPD	Link process definition	Definición de procesos de enlace.
PCDATA	Parsed character data	Caracteres de datos que pueden ser analizados en busca de marcado.
PI	Processing instruction	Instrucción de proceso.
RCDATA	Replaceable character data	Caracteres de datos en los que una referencia a una entidad puede ser reconocida y reemplazada.
SDATA	Specific character data	Datos que contienen caracteres dependientes del sistema.

SGML	Standard Generalized Markup Language	Lenguaje de marcado generalizado estándar.
XML	Extensible Markup Language	Lenguaje de marcado extensible.

C.2. Caracteres delimitadores en la sintaxis concreta de referencia

Término	**Significado**	**Carácter Asociado**	**Descripción**
ERO (&)	Entity reference open	&	Carácter con el que empezamos una referencia a una entidad general.
AND	And connector (within declaration group)	&	En un grupo, todos los elementos del grupo deben aparecer. Pueden aparecer en cualquier orden.
CRO	Character reference open	&#	Caracteres con los que empieza una referencia.
PERO (%)	Parameter entity reference open	%	Carácter con el que empezamos una referencia a una entidad de tipo paramétrica.
REFC (;)	Entity reference close	;	Carácter con el que terminamos la referencia a una entidad.
STAGO	Start-tag open	<	Carácter con el que empieza una etiqueta de inicio de elemento.
ETAGO	End-tag open	</	Caracteres con los que empieza una etiqueta de fin de elemento.
MDO (<!)	Markup declaration open	<!	Delimitador con el que comienza una declaración.
PIO	Processing instruction open	<?	Caracteres con los que empieza una instrucción de proceso.
TAGC	Tag close	>	Carácter de cierre de etiqueta
MDC (>)	Markup declaration close	>	Delimitador de fin de declaración.
PIC	Processing instruction close	>	Carácter con el que finaliza una instrucción de proceso.
GRPO	Group open	(	Abre un grupo.
GRPC	Group close	)	Cierra un grupo
DSO	Declaration subset open	[	Carácter con el que empieza un subconjunto de una declaración.
DSC (])	Declaration subset close	]	Carácter con el que finaliza un subconjunto de una declaración.

DTGC (])	Data tag group close	]	Carácter que abre un grupo DATATAG.
DTGO ([)	Data tag group open	[	Carácter que cierra un grupo DATATAG.
LIT	Start or end of literal string	"	Carácter con el que empiza y termina una cadena literal.
LITA	Alternative start or end of literal string	’	Carácter alternativo con el que puede empezar y terminar una cadena literal.
MSC	Marked section close	]	Carácter de fin de sección.
VI	Value indicator (within attributes)	=	Carácter usado para indicar el valor de un atributo.
COM	Start and end of comment	–	Caracteres que distinguen el inicio y final de un comentario.
MINUS	Exclusion set identifier	-	Carácter que distingue un grupo de exclusión (los grupos de exclusión suprimen elementos del modelo de contenido)
PLUS	Inclusion set identifier	+	Carácter que distingue un grupo de inclusión (los grupos de inclusión añaden elementos al modelo de contenido)
PLUS	Required and repeatable occurrence indicator	+	Indicador de ocurrencia. Indica que un elemento puede ocurrir una o más veces.
REP	Optional and repeatable occurrence indicator	*	Indicador de ocurrencia. Indica que un elemento puede ocurrir cualquier número de veces (incluso ninguna vez).
OPT	Optional occurrence indicator	?	Indicador de ocurrencia. Indica que un elemento puede ocurrir una o ninguna vez.
OR	Or connector (within declaration group)	|	Conector. En un grupo, sólo uno de los elementos que conecta debe aparecer.
SEQ	Sequence connector (within declaration group)	,	Conector. En un grupo, todos los elementos que conecta debe aparecer y deben hacerlo en el orden indicado.
NET (/)	Null end-tag	/	Carácter que puede sustituir una etiqueta de fin. Ver su uso en la sección SHORTTAG.
RNI	Reserved name indicator	#	Distingue una palabra reservada de un identificador genérico.

C.3. Delimitadores que pueden usarse como referencias cortas

Cadena	Número	Descripción
&#TAB;	9	Horizontal tab
&#RE;	13	Record end
&#RS;	10	Record start
&#RS;B	10,66	Record start followed by leading blanks
&#RS;&#RE;	10,13	Empty record
&#RS;B&#RE;	10,66,13	Blank records
B&#RE;	66,13	Trailing blank(s) followed by record end
&#SPACE;	32	Space
BB	66,66	Two or more blanks (spaces or tabs)
"	34	Quotation mark
#	35	Number sign
%	37	Percent
’	39	Apostrophe
(	40	Left parenthesis
)	41	Right parenthesis
*	42	Asterisk
+	43	Plus sign
,	44	Comma
-	45	Hyphen
–	45,45	Two hyphens
:	58	Colon
;	59	Semicolon
=	61	Equals sign
@	64	Commercial at
[	91	Left square bracket
]	93	Right square bracket
^	94	Circumflex accent
_	95	Low line
{	123	Left curly bracket
|	124	Vertical line
}	125	Right curly bracket
~	126	Tilde

Apéndice D

Declaración SGML del analizador OpenSP

Declaración SGML usada por defecto por el analizador OpenSP.

```
 <!SGML "ISO 8879:1986"
                              CHARSET
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
CAPACITY PUBLIC    "ISO 8879:1986//CAPACITY Reference//EN"
SCOPE    DOCUMENT
SYNTAX
SHUNCHAR CONTROLS 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
         18 19 20 21 22 23 24 25 26 27 28 29 30 31 127 255
BASESET  "ISO 646-1983//CHARSET International Reference Version
          (IRV)//ESC 2/5 4/0"
DESCSET  0 128 0
FUNCTION RE                    13
         RS                    10
         SPACE                 32
         TAB       SEPCHAR     9
NAMING   LCNMSTRT  ""
         UCNMSTRT  ""
         LCNMCHAR  "-."
         UCNMCHAR  "-."
         NAMECASE  GENERAL     YES
                   ENTITY      NO
```

```
DELIM    GENERAL   SGMLREF
         SHORTREF  SGMLREF
NAMES    SGMLREF
QUANTITY SGMLREF
         ATTCNT    99999999
         ATTSPLEN  99999999
         DTEMPLEN  24000
         ENTLVL    99999999
         GRPCNT    99999999
         GRPGTCNT  99999999
         GRPLVL    99999999
         LITLEN    24000
         NAMELEN   99999999
         PILEN     24000
         TAGLEN    99999999
         TAGLVL    99999999
                           FEATURES
MINIMIZE DATATAG   NO
         OMITTAG   YES
         RANK      YES
         SHORTTAG  YES
LINK     SIMPLE    YES 1000
         IMPLICIT  YES
         EXPLICIT  YES 1
OTHER    CONCUR    NO
         SUBDOC    YES 99999999
         FORMAL    YES
                          APPINFO NONE>
```