MODELO ESTRUCTURAL NORMALIZADO DE INSTRUMENTOS DE DESCRIPCIÓN DE ARCHIVO

Eduardo Peis y J. Carlos Fernández-Molina
Facultad de Biblioteconomía y Documentación
Universidad de Granada

Resumen:

La estructuración mediante lenguajes de etiquetado de instrumentos de descripción de archivos que permitan la difusión a nivel internacional de la información, es un paso clave hacia la necesaria adaptación de los archivos para el aprovechamiento de las posibilidades tecnológicas. Una herramienta adecuada para ello es un DTD SGML para instrumentos de descripción archivística denominado EAD (Encoded Archival Description). Se presenta una estrategia metodológica que, partiendo del análisis de la EAD y del objeto informativo a etiquetar, permite la creación semiautomática de una versión digital del mismo.

Palabras clave: Archivos / Información electrónica / SGML / EAD

1.- Introducción

La información es registrada en soportes cada vez más variados, con formatos cada vez más complejos y su volumen de crecimiento sigue una tendencia exponencial. Las posibilidades tecnológicas aplicadas al tratamiento y gestión de la información y los documentos han determinado la aparición de nuevos comportamientos en los usuarios, lo que ha de provocar que los servicios de información sean sometidos a revisión con el objetivo de hacer frente a todo tipo de materiales y formatos electrónicos y ofrecer un acceso integral a los recursos informativos, incluso externos. No se trata de partir de cero sino, como decía Croft (1) al referirse a las líneas de investigación actuales en Recuperación de la Información, de reorientar los esfuerzos.

En el contexto de los archivos, el mínimo desarrollo normativo, la escasez de recursos económicos y administrativos y sus especiales características, hacen que esta necesidad de reorientación sea aún más acuciante. En efecto, los archivos deben acomodarse a las nuevas posibilidades tecnológicas, ya que su concurso puede resultar fundamental para rentabilizar los recursos, incrementando el grado de explotación de la información. Un paso clave en este esfuerzo de adaptación es el diseño de instrumentos de descripción estructurados de tal forma que permitan la difusión a nivel internacional de la información de archivo.

Hay una serie de factores, relativos a la descripción archivística, que influyen en el desarrollo de los citados instrumentos de descripción: la adición de nuevo material a clases ya existentes, la variación en la profundidad o complejidad de la descripción, la posibilidad de una presentación parcial de la información, la diversidad de objetivos perseguidos, la variedad de puntos de acceso en un mismo sistema, la dificultad en la coordinación de sistemas, la variedad de terminología, la dificultad de relacionar las descripciones archivísticas con sistemas nacionales o internacionales de catalogación y normas bibliográficas, los diferentes códigos de referencia, el formato de presentación e intercambio o el potencial de los diferentes motores de recuperación.

La práctica profesional tradicional ha intentado solucionar algunos de estos problemas haciendo corresponder diferentes instrumentos de descripción con las diferentes agrupaciones documentales posibles, establecidas de acuerdo con los grandes principios de organización archivística. El resultado es un modelo generalizado en su concepción global, pero poco transferible en su aplicación particular, que además no resuelve muchas de las cuestiones apuntadas. Por lo tanto, es obligado diseñar tales sistemas teniendo en cuenta las siguientes

necesidades: presentar de forma extensiva e interrelacionada la información descriptiva contenida normalmente en los instrumentos de descripción, preservar las relaciones jerárquicas que existen entre niveles de descripción, representar información descriptiva que es "heredada" de un nivel jerárquico a otro, "navegar" en una arquitectura de información jerárquica y realizar indización y recuperación de elementos específicos.

Estas necesidades se pueden satisfacer con éxito creando una versión "digital" de dichos instrumentos de descripción, codificando la información descriptiva estructurada con un lenguaje de "etiquetado" normalizado internacionalmente.

En los sistemas bibliotecarios, para posibilitar la importación/exportación de información en formato electrónico y la creación de bases de datos que permitan la inclusión de información en múltiples formatos diferentes, se está empleando el lenguaje normalizado SGML (2). La norma ISO SGML (3) es un metalenguaje, qes decir, un medio de describir formalmente un lenguaje, en este caso, un lenguaje de codificación etiquetado. Es un sistema "descriptivo" que se sirve de códigos que simplemente ofrecen nombres para categorizar e identificar partes de un documento. Esto significa que SGML es un protocolo elaborado para expresar estructuras de contenido más que apariencia de documentos, es decir, usa códigos de marcaje (etiquetas) que simplemente proporcionan nombres para categorizar partes de un documento. Por tanto, identifica cada parte de un documento por la finalidad que tiene dentro del mismo utilizando lo que se denomina codificación descriptiva. Con el etiquetado descriptivo en lugar del de procedimiento, el mismo documento puede ser procesado fácilmente con muchos tipos diferentes de *software*, cada uno de los cuales puede aplicar diferentes instrucciones de procesamiento.

SGML introduce la noción de "tipo de documento" y, consiguientemente, un *document type definition* (DTD). El tipo de un documento es definido formalmente por sus partes constituyentes y su estructura. Esto implica, entre otras cosas, que diferentes documentos del mismo tipo pueden ser procesados de una manera uniforme. Por otra parte, SGML proporciona un mecanismo de aplicación general para la sustitución de cadenas (*string substitution*), que es una forma simple de asegurar la independencia de sistemas concretos (4).

De igual manera, debido a su potencia como estructurador del contenido, SGML facilita la accesibilidad al mejorar la discriminación informativa. La capacidad para incluir enlaces internos y externos a otros documentos maximiza las posibilidades de *browsing* y su independencia de los datos asegura una capacidad innata para la integración de todo tipo de objetos informativos.

No es de extrañar que el primer intento de aplicación de una codificación normalizada a los instrumentos de descripción archivística iniciado por la Universidad de California en Berkeley y dirigido por Daniel Pitti (5), seleccionase SGML (6) como técnica ideal para llevar a cabo dicha codificación.

El resultado de este proyecto fue el diseño de un DTD que define una clase de documentos que, en términos generales, constan de una página de título opcional, la descripción de una unidad de material archivístico y unos apéndices también opcionales. La página de título, que presentaba el borrador del DTD, podría incluir variados elementos como la identificación del fondo o el tipo de instrumento de descripción. Una unidad de descripción, de acuerdo con el DTD, podría incluir una breve descripción de la unidad (utilizando elementos etiquetables análogos a los empleados en un registro catalográfico MARC), una más amplia descripción narrativa de la unidad y cualesquiera partes segregables (incluyendo elementos etiquetables como título, fechas, alcance y contenido) y una lista formateada de las partes que contienen a dicha unidad.

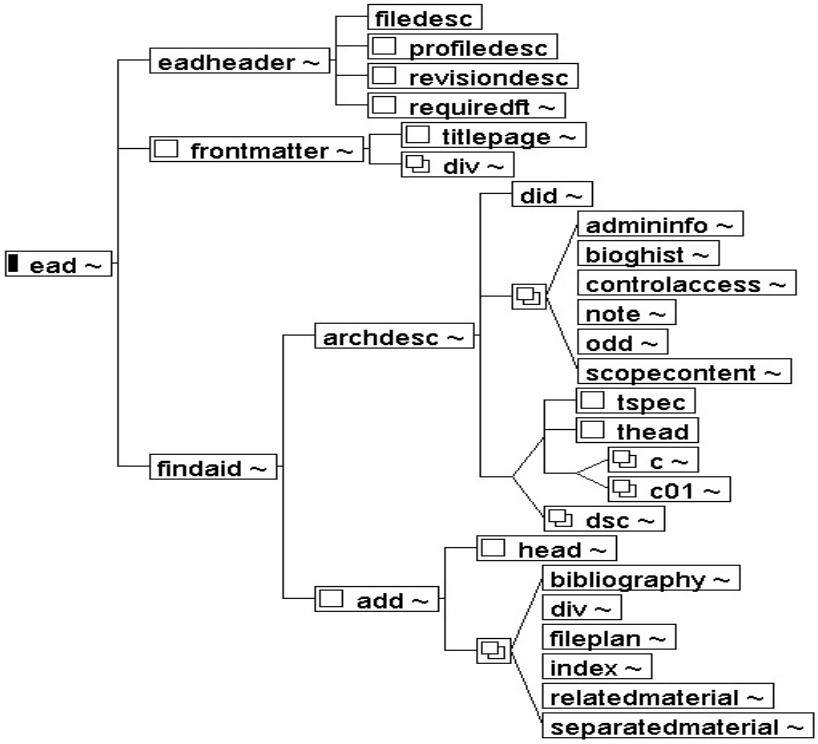
Incluso aceptando la premisa de las ventajas de aplicación de este modelo, muchos sistemas de archivo pueden encontrar la dificultad insalvable de "convertir" sus instrumentos de descripción en soporte papel a formato digital, adecuando además los componentes informativos de dichos instrumentos al modelo estructural EAD. En este trabajo, partiendo de un análisis del modelo SGML a emplear (EAD) y del objeto informativo a etiquetar, proponemos una estrategia metodológica que permitiría la creación semiautomática de una versión digital del instrumento de descripción.

2.- Metodología

Si tenemos en cuenta la variabilidad de situaciones con las que nos podemos encontrar en este tipo de servicios de información, es conveniente diseñar un método modular, lo que implica la distribución del desarrollo metodológico en diferentes fases: una primera fase de análisis previos, una segunda de captura/conversión y una tercera de organización.

La primera fase metodológica, y quizás la más importante, está constituida por los análisis iniciales tanto del DTD EAD, como del instrumento de descripción en sí mismo. Su objetivo es triple: comprobar la adecuación del modelo SGML al instrumento de descripción a etiquetar; estudiar la posible correspondencia estructural, concretando aquellos elementos informativos que no presenta el modelo "tradicional" y que tendrán que ser aportados de forma "manual"; y descubrir características físicas y lógicas del formato del documento "fuente" que permitirán proceder a etiquetarlo SGML de forma automática.

Para comprobar la adecuación del modelo es necesario conocer a fondo la EAD. El DTD (7) ha sido diseñado para reflejar la jerarquía natural que presenta la organización de los fondos, en conjunción con la jerarquía intelectual que imponen los archiveros con sus prácticas descriptivas. Contiene dos tipos de elementos: aquellos que codifican puntos específicos en la descripción de partes componentes del instrumento de descripción o el material que describe (elementos descriptivos); y aquellos que podrían codificar cualquier característica del documento (elementos genéricos). Estos últimos generalmente son incluidos en elementos descriptivos. Los principales componentes informativos que constituyen el modelo se pueden representar mediante una estructura arborea que refleja claramente las relaciones jerárquicas entre los elementos:



EAD usa el término instrumento de descripción (finding aid) para referirse a cualquier herramienta jerárquica que haya sido codificada usando EAD y que permitirá a un creador de registros o a un usuario acceder a los materiales que están siendo descritos.

A un nivel muy básico, un documento "instrumento de descripción" codificado utilizando EAD, consta de tres segmentos: uno que proporciona información sobre el instrumento de descripción en sí mismo (su título, compilador, fecha de compilación) (<eadheader>); un segundo componente que incluye las cuestiones preliminares necesarias para la publicación formal del instrumento de descripción (<frontmatter>); y un tercero que proporciona la descripción del material archivístico en sí misma, además de la información contextual y administrativa asociada (<findaid>).

El <eadheader>, que está basado en el elemento "header" del modelo SGML Text Encoding Initiative -TEI- (8), abarca cuatro subelementos (<filedesc>, <profiledesc>, <revisiondesc> y <requiredft>) para capturar o contener la mayoría de la información que

normalmente es registrada sobre la creación, publicación y la distribución de un documento "instrumento de descripción". Otra información adicional sobre el instrumento de descripción que no "cuadra" con el modelo TEI puede ser incluida en el elemento <frontmatter>, que incluida en dos elementos como <titlepage> y refleja detalles de tipo introductorio (prefacio, introducción, etc.) necesarios para la publicación formal del instrumento de descripción.

En el elemento <findaid> pueden presentarse dos tipos de información que corresponden a los dos subelementos principales que incluye. El elemento "descripción archivística" o <archdesc> presenta información jerárquicamente organizada que describe una unidad de registros o papeles junto con sus partes componentes o divisiones. El elemento "complemento a los datos descriptivos" o <add> contiene información suplementaria opcional que no describe directamente los registros pero facilita su uso por parte de los investigadores (por ejemplo, una bibliografía).

Cualquier atributo o valor que sea usado para describir materiales a un nivel es automáticamente "heredado" por los elementos a niveles subordinados, a menos que EAD especifique lo contrario.

Los atributos reflejan propiedades definidas de un elemento y pueden tomar diferentes valores, dependiendo del contexto en el que aparezcan. Para configurar uno o más atributos, un codificador debe incluir el nombre del atributo (utilizando los mismos signos "<" y ">" que se utilizan para definir etiquetas) junto al valor que tiene dicho atributo.

La conclusión, en términos generales, de esta primera fase fue que toda la información contenida en un instrumento de descripción "tradicional" puede ser estructurada utilizando los elementos que componen la EAD.

Para archivos que tienen sus instrumentos de descripción en formato de base de datos, el proceso de conversión a EAD no es demasiado difícil, aunque sí algo complejo. Es posible utilizar funciones *macro* de procesadores de texto o diseñar *script* de conversión que permitan relacionar el contenido de los elementos presentes en los registros de la base de datos con los elementos correspondientes del modelo SGML, insertando las etiquetas adecuadas. Aquellos instrumentos de descripción que no existen en formato electrónico, deben ser convertidos previamente a dicho formato, bien rehaciendo el instrumento de descripción o bien utilizando tecnología de reconocimiento óptico de caracteres (OCR).

No obstante, en ambos casos (formato de base de datos o soporte papel) las inconsistencias de formateo o incluso idiosincrasias descriptivas individuales hacen complicado el proceso de conversión. Es básica, por lo tanto, la segunda etapa de la primera fase: el análisis formal del objeto informativo a convertir. El hallazgo de características de formato uniformes (tabulaciones, espacios, columnas regulares, etc.) pueden permitir la identificación automática de los diferentes elementos informativos, posibilitando el etiquetado.

El siguiente paso, por lo tanto, es el proceso de captura/conversión. La captura mediante escáner necesariamente implica el procesado OCR. La adecuada configuración del *software* OCR permitirá "salvar" aquellas características de formato que serán fundamentales para la conversión posterior. Nuestra propuesta hace de esta segunda fase metodológica una etapa intermedia, ya que incluye la utilización de una aplicación de uso público para la conversión SGML como *Rainbow Maker* (9), que utiliza la información a nivel de párrafo, a nivel de carácter e incluso atributos comunes que no son dependientes del estilo, para insertar etiquetas (en este caso etiquetas *Rainbow*) que aislan e indican las cadenas de texto relevantes. Aunque *Rainbow* es un formato SGML (en concreto un DTD), no es apropiado para la representación permanente de los datos, ya que no contiene más estructura o identificación de contenido que la que se halla en el documento original.

Desarrollar este paso intermedio en lugar de utilizar directamente alguna facilidad de programación para la conversión, implica, a corto plazo, un gasto adicional en tiempo de procesamiento, pero a medio plazo representa un doble beneficio: por una parte contamos con un fichero SGML intermedio que puede ser empleado, aplicando rutinas de conversión inmediatas, para la generación de múltiples objetos informativos; y, por otra parte, permite utilizar características comunes que simplificarían la producción de especificaciones que pueden ser aplicadas a más de un instrumento de descripción, aumentando así el componente automático del proceso.

La última fase consiste en el diseño de una serie de sencillos *script* para la conversión y organización de las etiquetas SGML. Un último paso permitirá "conectar" todos los componentes informativos necesarios para obtener el fichero producto completo.

3.- Resultados

El proceso es semiautomático porque necesariamente implica la introducción manual, en un fichero denominado plantilla, de toda aquella información necesaria para el documento SGML que no está presente en el documento fuente. Por ejemplo, la información contenida en los elementos <eadheader> y <frontmatter>, así como las aclaraciones de sistema (content-type, content-ID y content-Description), la aclaración EAD y las referencias de entidad han sido aportadas manualmente, incluyéndolas en el fichero "plantilla".

La configuración inicial de los atributos necesarios ha sido realizada también de forma previa al procesamiento, presentando los correspondientes valores en las aclaraciones de atributos contenidas en la aclaración EAD.

En el caso de los atributos, naturalmente ha sido aprovechada la facilidad de empleo de los denominados valores por defecto, esto es, aquellos que son aportados automáticamente por el sistema si el codificador del instrumento de descripción no especifica un valor alternativo. Por ejemplo, en el elemento <ead> que sirve para indicar al ordenador el procesamiento de un documento codificado en EAD, hay un atributo llamado audience. Este atributo indica si los contenidos de este documento codificado pueden ser accesibles a todos (es decir, públicos -public-) o restringidos al personal (es decir, privados -private-). Si el codificador no etiqueta este atributo específicamente como private, el sistema considerará automáticamente el atributo como público.

Los puntos de acceso controlado (como encabezamientos de autoridad y encabezamientos de materia) pueden insertarse en cualquier sitio donde halla una <p> (etiqueta que introduce un párrafo de texto cualquiera).

Las notas (notes) se definen de acuerdo con su localización. Todos los elementos que podrían ser usados en gran cantidad de otros elementos son definidos genéricamente, pero su estado (es decir, requerido/no requerido, repetible/no repetible) es dependiente del elemento en el que esté incluido.

Esta información de configuración también ha sido incluida en el fichero "plantilla". El uso de este fichero "plantilla", que presenta un identificador (ID) que ha sido utilizado como puntero para la conexión de esta información con el producto de la conversión, facilita el tratamiento de los ficheros obtenidos y, al evitar redundancias, ahorra espacio de almacenamiento.

El producto de la captura y procesamiento OCR del instrumento de descripción en formato papel consta de una serie de ficheros. Cada uno de ellos contiene texto en formato *RTF* (Rich Text Format) y conserva información de la existencia de fuentes en negrita y cursiva, además de una representación precisa de la presentación de página (incluyendo tabulaciones y líneas en blanco).

Rainbow Maker utiliza la información de formato y lógica de los documentos procesados contenidos en dichos ficheros para insertar las etiquetas de la DTD *Rainbow* y el producto se incluye en un único fichero.

Empleando facilidades del lenguaje de programación *perl* (*Practical Extraction and Report Language*) (10) se procede a la traducción "objeto a objeto" de etiquetas *Rainbow* a etiquetas EAD y a la organización de dichos objetos, conforme al modelo de documento adecuado. Una vez salvado, el fichero producto de esta operación es conectado a los datos esenciales contenidos en el fichero "plantilla" produciendo también un único fichero.

El modelo obtenido incluye todas las aclaraciones de elementos, atributos y referencias que debe presentar un DTD EAD.

4.- Conclusiones

La codificación SGML de instrumentos de descripción archivística accesibles localmente o en línea a través de las redes simplifica, mejora y expande el acceso a las colecciones archivísticas, haciendo posible la conexión registros catalográficos-instrumentos de descripción. Permite también la búsqueda en conjuntos de instrumentos de descripción conectados, lo que posibilita el acceso mediante palabras clave a fondos localizados y la recuperación de elementos informativos que de otro modo permanecerían ocultos.

EAD constituye un DTD SGML que es ideal para representar la estructura de los instrumentos de descripción archivística, ya que no sólo describe las partes físicas e intelectuales constituyentes de dichos documentos en términos de distintos campos o elementos, sino que también posibilita el mantenimiento de las relaciones jerárquicas entre dichos elementos, permitiendo así la "navegación" entre niveles de descripción y evitando la duplicación de información. Por otra parte, dado su carácter digital y normativo, resuelve muchas de las cuestiones apuntadas en el primer apartado de este trabajo. Su perfecta correspondencia con normas internacionales para la descripción archivística como la ISAD (International Standard Archival Description) (11) y otras normas y formatos, como el MARC, aseguran su versatilidad.

Es posible desarrollar un método que permite la codificación EAD de instrumentos de descripción ya existentes en soporte papel de forma semiautomática, usando la tecnología disponible hoy en día.

La utilización de SGML como herramienta clave en el método propuesto, hace que productos intermedios en la globalidad del proceso puedan ser utilizables por sí mismos, permitiendo de esta forma maximizar la explotación de los recursos tratados. La fácil manipulación de este fichero SGML y la versatilidad de las posibles rutinas de indización utilizables facilitan el empleo de métodos avanzados de recuperación. En este sentido, la naturaleza modular de la metodología empleada y el uso de lenguajes normalizados para su desarrollo, hace posible la integración de contenidos informativos procedentes de fuentes externas y permite la exportación completamente automatizada de información local en un formato estándar, además de posibilitar la reutilización de los productos intermedios en otros desarrollos, incluso externos.

5.- Referencias

[1] Croft, W. B. What do people want from information retrieval?: (The top 10 research issues for companies that use and sell IR systems). *D-Lib Magazine*, 1995

[2] http://www.dlib.org/dlib/november95/11croft.html (visitado el 24 de febrero de 1999)

[3] Corthouts, J., and R. Philips. SGML: a librarian's perception. *The Electronic Library*, 14 (2), 1996, 101-110.

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

[5] Sperberg-Mc Queen, C. M., and Burnard, L. A gentle introduction to SGML. http://www-tei.uic.edu/orgs/tei/sgml/teip3sg/SG.htm (visitado el 24 de febrero de 1999)

[6] Development of the Encoded Archival Description Document Type Definition. http://www.loc.gov/ead/eadback.html (visitado el 24 de febrero de 1999)

[7] Cover, R. The SGML/XML Web page. http://www.oasis-public.com/cover (visitado el 24 de febrero de 1999)

[8] Gilliland-Swetland, A. J. Encoded Archival Description Document Type Definition (DTD). Applications GuidelinesTechnical Document No. 1., 1996. http://www.loc.gov/ead/ (visitado el 24 de febrero de 1999)

[9] TEI Guidelines for Electronic Text Encoding and Interchange (P3) http://etext.virginia.edu/TEI.html (visitado el 24 de febrero de 1999)

2. Sklar, D. Accelerating conversion to SGML via the Rainbow format. *Electronic Book Technologies EBT*. http://ftp.sunet.se/pub/text-processing/sgml/Rainbow/Rainbow.why (visitado el 24 de febrero de 1999)

3. Practical Extraction and Report Language (PERL). http://www.bme.unc.edu/facilities/software/perl/perlIndex.html (visitado el 24 de febrero de 1999)

4. International Council on Archives. *ISAD(G): General International Standard Archival Description, adopted by the Ad Hoc Commission on Descriptive Standards, Stockholm, Sweden, 21-23 January 1993* (Ottawa, Ont.: International Council on Archives, 1994). http://www.archives.ca/ica/cds/isad(g)e.wp (visitado el 24 de febrero de 1999)