¿Qué enseñamos después del MARC?

Virginia Ortiz-Repiso Jiménez
Universidad Carlos III de Madrid

0.1. Resumen

Se analizan los sistemas de metadatos que existen en diferentes colectivos para solucionar los problemas que plantea la búsqueda y recuperación efectiva de recursos en Internet. Se comparan, además, los nuevos sistemas con los registros bibliográficos utilizados en las bibliotecas. Por último, se estudia la posibilidad de que estos sistemas puedan remplazar las herramientas que se utilizan de forma generalizada en las bibliotecas : el formato MARC y las Reglas de Catalogación (Autor).

Palabras clave: Metadatos. Formato MARC. SGML.

0.2. Abstract

Metadata systems to solve searching and retrieval Internet problems are analysed. New metadata systems and traditional bibiographic records are compared. At last, the posibility that those systems may be replace the tradicional library tools, such MARC format and Cataloguing Rules, are examined (Author).

Keywords: Metadata. MARC. SGML.

1. Introducción

El título de esta comunicación es *¿Qué enseñamos después del MARC?*, y voy a centrarme principalmente en el impacto que, en el mundo bibliotecario en general y en el proceso catalográfico en particular, están teniendo los nuevos sistemas y estándares que se utilizan para describir, recuperar y acceder a los objetos de información en un entorno de red.

2. Internet y las herramientas de búsqueda

El rápido desarrollo del World Wide Web y el incremento de los recursos disponibles en Internet ha hecho necesaria la creación de herramientas que proporcionen acceso a los millones de documentos que existen en formato electrónico.

La Red, a través de las máquinas de búsqueda, utiliza métodos automáticos para identificar los recursos de Internet. Existe una gran cantidad de programas que navegan automáticamente a través de los espacios Web, buscando enlaces, recuperando documentos, indizándolos y creando bases de datos con ellos. Ahora bien, estos sistemas de gran potencia, recuperan gran cantidad de documentos pero con muy poca precisión. La causa principal no es que los métodos automáticos que utilizan describan de forma poco adecuada los recursos en la Red sino que los propios documentos HTML carecen de datos suficientes de descripción del recurso. Como resultado hay mucho ruido a la hora de la recuperación y muchas veces se convierten en inaccesibles para el más paciente de los buscadores en la red.

Además, mientras que el tamaño de Internet crezca exponencialmente será cada vez más difícil moverse por esta "masa" indiscriminada de resultados de búsqueda. Al mismo tiempo, se corre el peligro de que las bases de datos creadas por estos mecanismos puedan llegar a ser mayores incluso, que los propios recursos de Internet.

A medida que la Red crece y que estos métodos de descripción se muestran cada vez más inadecuados, una amplio colectivo, que incluye profesionales de la información, informáticos de redes, diseñadores de *software*, investigadores y un amplio etcétera, ha empezado a plantearse la necesidad de crear descripciones y catálogos que identifiquen los recursos electrónicos en Internet de una manera más eficaz y permitan una búsqueda y recuperación más efectiva.

Encontrar una solución a este problema se ha convertido en materia de estudio para muchas instituciones y asociaciones como pueden ser: la Biblioteca del Congreso, OCLC, la American Libray Association, la National Science Foundation, etc.

Todas las personas y colectivos involucrados en los problemas que plantea la búsqueda y recuperación de recursos en las redes coinciden en que la mejor manera de solucionar el problema es la creación de metadatos que describan los recursos.

2. Metadatos para la recuperación de la información electrónica

El término *metadata* o metadatos se está utilizando cada vez más para referirse a la disponibilidad de datos sobre los recursos de información. Los metadatos son un conjunto de datos que pueden usarse para describir y representar objetos de información. Contienen un conjunto de elementos de datos que pueden usarse para describir el contenido y la localización de un objeto de información y facilitar su recuperación y acceso en un entorno de red. En otras palabras, son

datos a cerca de datos. Los metadatos en sí no son algo nuevo. Lo que es nuevo hoy en día es la multitud de métodos que se están creando y la forma de usarlos.

Los registros bibliográficos que se han creado a lo largo de muchos años en el mundo bibliotecario son esencialmente metadatos. Proporcionan información descriptiva y analítica sobre un objeto de información. Los catalogadores los han empleado como método descriptivo desde hace décadas, bien como registros MARC en los OPACs, bien como fichas catalográficas en los catálogos manuales. Un registro catalográfico no es otra cosa que un conjunto de metadatos. Nosotros podríamos llamarlo catalogación pero para algunas personas este término conlleva una carga excesiva: formatos MARC, Reglas de Catalogación, etc. El término metadatos se utiliza pues como un término neutral (Caplan, 1995).

La definición más exacta del término en el contexto Internet es la realizada por Tim Berners-Lee (1996): información legible por ordenador sobre recursos web. La frase *legible por ordenador* es clave. Hablamos de información que los diferentes *softwares* pueden utilizar para hacernos más fácil la tarea de búsqueda y recuperación de recursos en la red. El mundo del World Wide Web es un mundo de información y parte de esa información es información sobre información. Ésta puede localizarse en el propio documento o puede formar una entidad separada, o puede transferirse acompañando al documento.

Los metadatos son importantes en la recuperación de la información global en Internet por distintas razones:

- Permiten indizar gran cantidad de datos de diferentes tipos sin necesidad de utilizar un gran ancho de banda ya que se indiza la representación del objeto y no el objeto en sí.
- Ayudan a descubrir y recuperar recursos en la red ya que analizan el contenido del objeto en profundidad.
- Comparten e integran recursos de información heterogéneos y localizados en sitios muy diversos.
- Pueden controlar el acceso a información restringida.

El uso de metadatos para organizar el contenido de la información en Internet se está extendiendo cada vez más. Existen, básicamente, tres formas de crearlos: por el autor cuando éste crea su recurso (elemento META en HTML, cabeceras SGML); por medios automáticos (SOIF-Summary Object Interchange Format- en Harvest); o por un servicio de información. También se puede crear con ellos una base de datos, central o distribuida, con punteros a los recursos que describen. Además, han evolucionado desde formatos de estructura muy simple a formatos más complejos. Y se han movido desde estándares emergentes pro-

pietarios a estándares internacionales. Y, muy importante, los metadatos que se crean se pueden compartir con otros (1).

La información que contienen es variada: desde información descriptiva similar a la que estamos acostumbrados a ver en las bibliotecas, hasta información que ayude a la aplicación cliente a tomar una decisión sobre el formato o sobre la localización. Sus usuarios pertenecen también a distintas categorías, desde aquellos que desean conocer sólo los términos de disponibilidad de un recurso a aquellos que desean tener más información sobre el contenido del objeto informativo. Los recursos son, además, de distinta tipología: algunos tienen una existencia efímera y sólo necesitan una descripción somera; unos pueden ser simples, otros más complejos.

2.1. ¿Qué estándares siguen los metadatos?

Las bibliotecas, hasta el momento, parecían ser las únicas instituciones que contaban con una sintaxis ampliamente aceptada y bien regulada de creación de metadatos (MARC, Reglas de Catalogación, etc.). Pero, sin embargo, en los últimos años, con la extensión y expansión de recursos electrónicos en Internet, están emergiendo otras sintaxis, que proporcionan también metadatos, y que prometen una mayor funcionalidad.

Todos los conjuntos de metadatos que se están empleando para describir los recursos en las redes, bien sea como parte de los documentos, bien como entidades individuales pero enlazadas a los documentos, siguen la norma SGML —*Standard Generalized Markup Language* (Lenguaje de marcas estándar generalizado)— , estándar internacional desde 1986 (ISO 8879).

Aunque SGML es un estándar generalizado, no es, realmente, un lenguaje de marcas como tal. SGML no proporciona por sí sólo un lenguaje de marcas que uno puede simplemente llevarse a casa y aplicarlo a una carta, una novela, un artículo o un registro catalográfico. SGML es conocido como un metalenguaje, esto significa que no es un único lenguaje sino una norma amplia para construir lenguajes de marcas. SGML proporciona una sintaxis para definir y expresar la estructura lógica de los documentos, así como las convenciones para nombrar los componentes o elementos de los documentos. Se puede decir que SGML es un conjunto de reglas formales para definir lenguajes de marcas específicos para tipos específicos de documentos. Este lenguaje de marcas específico se denomina *Definición del Tipo de Documento* (DTD). Por ejemplo, la Asociación de Editores Americanos junto con OCLC ha desarrollado un juego de 3 DTDs: uno para libros, otro para publicaciones periódicas y otro para artículos de revista. HTML, tan conocido por todos, es un DTD específico de SGML. Puede haber, y de hecho ya existen, DTDs para bibliotecas, para museos y para archivos. No hace mucho el profesor Larson (1996) ha creado en Berkeley un DTD para el

USMARC con el propósito de utilizarlo en un prototipo de catálogo bibliográfico que emplea tecnología de recuperación avanzada.

Además, los DTDs que se comparten y se utilizan en una comunidad específica se pueden convertir y de hecho se convierten en estándares internacionales. Es el caso, por ejemplo, de la Asociación Americana de Editores con la norma ISO 12083.

La combinación de marcas descriptivas y DTDs permite que múltiples tipos de *software* puedan procesar documentos codificados en SGML. Es decir, SGML es independiente del *hardware* y *software* que se utilice. La información que se crea utilizando esta norma no se vuelve obsoleta si el *software* se queda anticuado o se cambia de programa. En esto se parece al formato MARC pero con una diferencia muy clara. La sintaxis SGML y sus reglas son precisas y es posible diseñar *software* que pueda ajustarse a cualquier DTD. Normalmente el *software* tiene un conjunto de herramientas que permite al usuario adaptar su funcionalidad a su DTD. Como resultado, el mercado de *software* SGML puede ser, en principio, cualquiera. De esta forma SGML está siendo utilizado por muchos productores de *software* de muy distintos tipos. Desde empresas que diseñan *software* específico para tratamiento de texto, como Word Perfect o Word de Microsoft, hasta empresas que se dedican a diseñar *software* para el sector de automóviles.

2.2. MARC versus SGML

Se ha señalado anteriormente que un registro bibliográfico es un tipo de metadatos. Difiere de otros en que desde hace más de tres décadas usa las Reglas de catalogación para modelar los datos, el formato MARC como esquema de codificación y, además, utiliza sistemas en línea propietarios para la recuperación de la información. La mayoría de los sistemas de metadatos en Internet no tienen, evidentemente, la misma estabilidad, ni unas normas tan bien reguladas, ni un esquema de codificación, ni sistemas específicos para recuperar información. Pero esto está cambiando y tenemos que examinar el impacto que está teniendo en el mundo bibliotecario.

El formato MARC es hoy en día un conjunto complejo de estándares para describir, almacenar, manipular y recuperar datos bibliográficos legibles por ordenador. Es un estándar altamente desarrollado que se diseño originalmente en los 60 para la descripción de libros impresos y que ha seguido adaptándose para proporcionar descripción, acceso y localización de la información de los recursos en la Red. En este sentido, el campo 856, recientemente incorporado (aunque no se utiliza en nuestro país) permite una descripción y enlace entre el registro MARC y el recurso electrónico que describe.

El formato MARC que ha sentado las bases de cooperación y comunicación de información bibliográfica en el mundo bibliotecario tiene, sin embargo, una serie de características negativas que podrían sintetizarse de la siguiente forma:

- Está estrictamente controlado, cualquier cambio o adición al formato tarda años en realizarse. Por ejemplo la tilde de la 'ñ' que tanto se utiliza en las URLs de las páginas personales ha tardado más de dos años en incorporarse al juego de caracteres del USMARC. EL proyecto Intercat de OCLC no podía operar con URLs que tuvieran ese carácter.
- Es laborioso, lento y costoso de realizar ya que debido a su complejidad debe realizarse por profesionales cualificados.
- Aunque está compuesto por campos de longitud variable, está limitado a una longitud máxima de 100.000 caracteres que, si bien es más que suficiente para los registros bibliográficos tradicionales de las bibliotecas, no lo es para otro tipo de registros de otros ámbitos (archivos, por ejemplo).
- Se adapta muy mal a información estructurada jerárquicamente ya que se basa en una estructura plana. Se diseñó para describir y acceder a la información de un registro bibliográfico y no para establecer relaciones entre registros.

Ante estas características, SGML proporciona un marco de trabajo prometedor por diversas razones:

- Puede tratar información jerárquicamente interrelacionada en tantos niveles como se necesiten.
- No tiene una limitación en el tamaño de los documentos.
- Es un estándar internacional adoptado por un número creciente de instituciones gubernamentales, de investigación y de la industria.

Tanto SGML como MARC son lenguajes de marcas descriptivos que crean textos estructurados. Pero SGML permite una flexibilidad máxima en el uso del texto. El usuario puede controlar los formatos de indización, presentación e impresión. Su estructura facilita la creación de bases de datos más sofisticadas que las basadas en MARC y permite la indización y recuperación del documento como documento y componentes del documento mediante búsqueda booleana, adyacencia, proximidad, ranking de relevancia, etc. La estructura del texto SGML soporta además navegación en línea avanzada.

Los registros pueden estar interrelacionados en distintos ficheros: el guión de una película y el vídeo, un mapa y el CD-ROM que lo contiene, por ejemplo.

En definitiva la flexibilidad del texto estructurado en SGML es incuestionablemente superior a la jerarquía plana del texto estructurado en MARC.

3. Conclusiones

Las estructuras de metadatos están adquiriendo un lugar central en la descripción de documentos electrónicos, de cualquier tipo y naturaleza, como medio de dotarlos de formas eficaces de recuperación.

El estándar que más se utiliza para la estructuración de estos datos es SGML que proporciona acceso a información jerárquica, bibliográfica y analítica compleja.

El formato MARC se ha mantenido fundamentalmente como una versión electrónica del catálogo en fichas con las limitaciones propias de este modelo. La función principal de este formato, cuando se diseñó por primera vez en los años 60, fue permitir la distribución electrónica de registros bibliográficos para la producción de fichas en papel. Al formato MARC le falta, en definitiva, habilidad suficiente para tratar información estructurada jerárquicamente y proporcionar acceso a colecciones complejas descendiendo por los niveles de análisis.

SGML se puede utilizar para crear una gran variedad de documentos y usarlo, además, para proporcionar acceso y control a multitud de formatos de información en línea. No es un lenguaje de marcas propietario como es el formato MARC y no depende tampoco de un *software* propietario o de una oferta limitada de *software*. En esencia podría ser posible usar SGML como el estándar general para las herramientas informativas que se usan para catalogar (Reglas, formato MARC, CDU, etc.), para crear los registros catalográficos y para crear los textos electrónicos de nuestro catálogo. Y, además, podríamos utilizar el texto para proporcionar acceso y control a los objetos digitales que no son texto.

Claramente hay ahora una oportunidad de crear y construir un entorno de información integrada en el que el catálogo proporcione acceso tanto a los documentos tradicionales como a información electrónica. Este entorno civilizado emergería de los desarrollos presentes en Internet y del tipo de proyectos que se están llevando a cabo.

Un OPAC podría servir de pasarela para acceder a una base de datos de metadatos en Internet, así como a las bases de datos de la propia biblioteca. A través de las capacidades de búsqueda del Z39.50 y las conexiones TCP/IP se podría, con un único interfaz, proporcionar acceso a los usuarios a bases de datos catalográfica, bases de datos en CD-ROM, bases de datos en texto completo, servicios en línea y otras bases de datos compatibles con el Z39.50. Los navegadores Web se convertirían así en clientes de los servidores Z39.50. Como resultado las bases de datos de los OPACs estarían integradas con una gran variedad de recursos internos de Internet, así como recursos externos (el Web de la Universidad de Yale es un ejemplo de ésto). Si así fuera se tendrían que estable-

cer perfiles Z39.50 para cada uno de los formatos de metadatos. Por el momento sólo cuentan con él el MARC y el Dublin Core.

Algunos autores como Amanda Xu (1997) mantienen que el formato MARC sigue siendo el más apropiado por su estructura y, sobre todo, porque su uso está muy extendido. Proponen crear programas de conversión entre los distintos formatos de metadatos con el objetivo de poder incluir los registros en los catálogos bibliotecarios. Yo no considero que el MARC sea la mejor forma de describir metadatos o que sea interesante convertir los metadatos en otros formatos a MARC. Pienso que podemos encontrarnos en la misma situación que cuando se instalaba un sistema integrado de gestión bibliotecaria y se seguía manteniendo el catálogo manual. Llegará un momento en que sólo uno de los dos sistemas prevalezca.

Ahora bien, en lo que coinciden la mayoría de los autores es en señalar que tanto las Reglas de Catalogación como los formatos MARC deben evolucionar para adaptarse al nuevo entorno.

El papel que jueguen los bibliotecarios en proporcionar descripciones y acceso a los recursos en las redes tendrá cada vez más relación con la creación, mantenimiento y mejora de descripciones de metadatos. En definitiva, buscar y estudiar metadatos y estándares para el futuro digital.

Los catalogadores deben adoptar nuevas perspectivas, nuevos métodos de organizar y recuperar la información. Deben, desde mi punto de vista, revisar los conceptos que van unidos a la forma de entender las relaciones que se establecen entre acceso, descripción y recuperación

El catálogo futuro debe ser flexible y multitarea y el MARC no puede proporcionarnos esa flexibilidad. Esto no quiere decir que el formato MARC ya no sirva. Hay billones y billones de registros MARC, el coste y el tiempo de convertirlos a SGML sería abrumador. Pero yo creo que el formato MARC no será por mucho más tiempo el único formato de codificación de datos bibliográficos en los sistemas bibliotecarios. El formato MARC no es suficientemente flexible para permitir a las bibliotecas aprovechar la tecnología de nuevo desarrollo de recuperación y acceso a la información.

4. Notas

(1) Se puede encontrar información muy completa sobre los distintos sistemas que existen, así como sus descripciones en DESIRE (Development of a European Service for Information on Research and Education: URL=<http://www.ukoln.ac.uk/metadata/DESIRE>

5. Referencias

Caplan, Priscilla (1995). You Call It Corn, We Call It Syntax-Independent Metadata For Document-Like Objects. // The Public Access Computer Systems Review. 6 : 4 (1995).

Berners-Lee, Tim. Metadata Architecture. URL= <http://www.w3.org/Design Issues/Metadata.html>

Larson, Ray. A few words about the USMARC.DTD. URL= <ftp://library.berkeley.edu/pub/sgml/marcdtd/README>

Xu, Amanda (1997). Metadata conversion and Library Opac. // Serial Librarian. (1997).