Metadatos y Catalogación ¿Conflicto cultural? Una experiencia con la catalogación de tesis

MICHAEL KREYCHE
Kent State University, E. U.

El término "metadatos" se encuentra en varios contextos con diversos sentidos. En un extremo la derivación del griego, paralela a la de la palabra "metafísica", nos explica que se trata de "datos sobre datos". En el otro, sospechamos que una rápida sustitución de "catalogación" o "datos" por "metadatos" no tenga otro motivo sino hacerlo parecer a uno muy al día.

Puesto que estos dos términos se refieren básicamente a una misma actividad, surge la pregunta de si debemos mantener una distinción entre ellas. Si aceptamos que el desarrollo de varios esquemas de metadatos ha profundizado las perspectivas de la profesión y nos ha dotado de nuevas capacidades, ¿no deberíamos considerar que la disciplina de catalogación abarca estas nuevas tecnologías y que resulta más amplia y rigurosa? Sin embargo la distinción persiste; ninguno de los términos ha reemplazado al otro.

Examinemos algunas de las características que tienden a diferenciar a estos dos conceptos. En primer lugar los metadatos se basan en una reexaminación de los elementos que constituyen una descripción de objetos y de colecciones de objetos. Este esfuerzo ha producido nuevos esquemas como el Núcleo de Dublín (en inglés, Dublin Core o DC), de utilidad general y muy sencillo, y otros más especializados y complejos, como la EAD, en inglés, Encoded Archival Description o EAD, que aunque se produjo como Metadatos para la descripción archivística, se siguió usando la sigla en inglés EAD.

Otra característica relacionada es que estos esquemas no son compatibles con los sistemas o catálogos tradicionales, diseñados para una u otra variación de MARC. Una nueva generación de un nuevo sistema, generalmente llamado repositorio, ha sido desarrollada para alojar a los nuevos formatos (pero podemos esperar que la próxima generación de sistemas aloje múltiples esquemas). La mayoría de estos sistemas han sido producidos por el ambiente académico, no por el sector comercial, y requieren cierto nivel de conocimiento técnico para su instalación y mantenimiento. Por lo tanto los profesionales que trabajan con estos sistemas pueden ser o considerarse tecnólogos más que bibliotecólogos. De hecho, es posible que no trabajen en bibliotecas sino en departamentos académicos o centros de investigación o computación.

Otra característica es que los nuevos esquemas de metadatos frecuentemente se apliquen a materiales inéditos o únicos coleccionados por bibliotecas, pero que hasta ahora han sido difíciles de organizar y manejar. Ejemplos de esta clase de recursos son objetos como fotografías, carteles, grabaciones de sonido o video, y trabajos escritos de producción académica como tesis e informes de investigación. Otro impulso importante para la organización de estos recursos es que muchos son de origen digital o que, fácilmente pueden convertirse en medios digitales.

Estos recursos que anteriormente yacían relativamente desatendidos en nuestras bibliotecas se benefician con la creación de repositorios que los organizan a lo cual hay que añadir la sencillez de los nuevos esquemas como el Núcleo Dublín, que sólo tiene quince elementos básicos. La resultante descripción, abreviada en comparación con la catalogación tradicional, representa una solución práctica que agiliza el procesamiento de estas colecciones.

Finalmente esta misma sencillez permite que la catalogación se haga no por catalogadores profesionales sino por especialistas en la temática del material –artistas, historiadores, científicos– o bibliotecólogos especializados en estas disciplinas y no en la catalogación. En el caso de tesis y otras obras académicas, el papel del catalogador puede ser asumido por la máxima autoridad: el autor mismo.

Algunas de estas diferencias se basan en la tecnología, pero dos de ellas tienen que ver con grupos distintos de personas: quienes manejan los sistemas y quienes se encargan de la "catalogación." Algunos miembros de cada uno de estos grupos pueden estar trabajando en la comunidad de metadatos o en la de catalogación, por lo que no es sorprendente que surjan choques entre estas dos comunidades, y si analizamos la distinción como fenómeno cultural, tal vez esto nos ayude a entender esos conflictos y a evitar que dañen la profesión y, por extensión, a los usuarios de nuestros servicios.

Examinemos el caso OhioLINK, un consorcio de bibliotecas académicas que mantiene una serie de sistemas al servicio de unas 90 instituciones, principalmente estatales. El primer servicio instalado hace unos 25 años, fue una red de sistemas integrados que alimentan con registros bibliográficos a un catálogo central por medio del cual los usuarios de cualquier institución pueden iniciar préstamos de cualquier otra.

Además, el consorcio coordina suscripciones que incluyen 140 bases de datos referenciales, 12,000 revistas electrónicas y 25,000 libros electrónicos. La mayoría de estas revistas y libros están en sistemas construidos y mantenidos por OhioLINK. Existen también repositorios para medios electrónicos (imágenes y grabaciones de sonido y video) y para tesis. Un metabuscador facilita la recuperación de recursos de más de uno de estos sistemas, y un servidor de resolución de enlaces conecta las citas de las bases de datos referenciales con los textos completos. En los primeros años estos sistemas generalmente fueron desarrollados con base en software comercial, pero la tendencia en los últimos años ha sido utilizar software gratuito de código abierto. Este es el caso con el repositorio llamado "Electronic Thesis and Dissertation Center" (o Centro de Tesis Electrónicas), el cual vamos a estudiar en más detalle.

El Centro de Tesis Electrónicas comprende una solución para la entrega, la revisión, la aprobación, el descubrimiento, y la distribución de tesis de las instituciones participantes. El sistema fue construido sobre una base de datos relacional integrada con un servidor web que genera el contenido de las páginas de manera dinámica. La interfaz web tiene tres componentes que atienden o sirven cinco fun-

ciones. El primer componente es el administrativo, que les permite a los estudiantes entregar la versión final de su tesis. La entrega consiste no sólo en subir la copia electrónica de la tesis sino también en ingresar algunos detalles personales y los metadatos del trabajo. No se previó la necesidad de que intervinieran catalogadores; el autor mismo asumiría ese papel. El siguiente paso administrativo es una revisión que hace el decanato. Si todo está en orden, el trabajo recibe la aprobación y se publica la tesis.

La publicación significa que la tesis tiene acceso público mediante los otros componentes web, un motor de búsqueda propio del Centro de Tesis Electrónicas y una implementación del Protocolo de la Iniciativa para Cosechar Metadatos en los Archivos Abiertos (en inglés, Open Archives Initiative–Protocol for Metadata Harvesting[1] o OAI-PMH). Este Protocolo utiliza el transporte HTTP Hyper Text Metadata Language (Lenguaje de Metadatos para Hipertexto), pero el contenido de las respuestas no consiste en páginas de HTML Hyper Text Markup Languaje, (Lenguaje de Marxcado para Hipertexto) sino en mensajes XML Extensible Markup Language (Lenguaje Generalizado de Marcado).

Este Protocolo fue desarrollado para facilitar la diseminación de contenidos dentro del contexto de la producción académica, especialmente en apoyo al movimiento de Acceso Libre o Acceso Abierto (en inglés, Open Access), que se define como "el acceso libre, inmediato, e irrestricto a material digital educativo y académico, principalmente artículos de investigación científica de revistas especializadas."[2]

El Protocolo define un mecanismo de comunicación entre dos sistemas, uno de los cuales es el repositorio el cual es denominado el "proveedor de datos". A otro se le llama "proveedor de servicios" y este último puede copiar, indizar, o manipular de otra manera los datos recuperados de uno o más repositorios para ofrecer un servicio determinado, por ejemplo un motor de búsqueda. El protocolo sólo

1 Open Archives Iniciative. http://www.openarchives.org/ (consultado el 29 de septiembre, 2008).

2 Acceso libre, http://es.wikipedia.org/w/index.php?title=Acceso_libre&oldid=20266122 (consultado el 28 de septiembre, 2008).

rige la comunicación entre los dos sistemas y no tiene relación alguna con la presentación del servicio al usuario.

Este protocolo es muy sencillo y cuenta con sólo seis "verbos" o mandatos. Uno es para solicitar la identificación del proveedor de datos; otro para solicitar un listado de esquemas de metadatos disponibles, y otro para solicitar un listado de los nombres de colecciones que existen dentro del repositorio. Hay dos verbos más para recuperar listados de registros en masa, uno para registros breves y otro para registros completos. En ambos casos los únicos criterios posibles para limitar el alcance de la solicitud son por colección y por fecha. No se puede buscar por autor, título, tema, etcétera. El último verbo es para recuperar un solo registro por su identificador único.

El Centro de Tesis Electrónicas de OhioLINK comenzó a funcionar en el 2001 con la participación de seis instituciones, pero gradualmente se unieron al proyecto otras instituciones y actualmente participan 20 con un total de más de 16,000 tesis. La Universidad Kent State inició su participación en 2004 con tesis de doctorado y en 2006 optó por incluir tesis de maestría también; hasta la fecha ha contribuido con 360 documentos para el repositorio.

Como algunas otras instituciones, la biblioteca de la Kent State University ha tenido durante muchos años la política de catalogar todas las tesis emitidas por la universidad y de pasar sus registros MARC Machine Readable Catalogs, (Catálogo Legible por Máquina) a OCLC (Online Computer Library Center, Centro Bibliotecario de Computación en Línea). Para continuar esta tradición con las tesis electrónicas los catalogadores comenzaron a estudiar la aplicación de normas de catalogación a este tipo de documento y participaron con el comité de bibliotecarios de procesos técnicos de OhioLINK en la preparación de pautas para todo el consorcio.[3] Al mismo tiempo se plantea-

3 Database Management and Standards Committee. Standards for Cataloging Electronic Theses and Dissertations – Remote Electronic Version (non-Reproduction), 4/2/2007 (Comité de Administración de Bases de Datos y estándares. Estándares para catalogar tesis y disertaciones electrónicas-Versión electrónica remota [no reproducción], 4/2/2007.

ron varias ideas como aprovechar los metadatos almacenados en el repositorio.

Resultaba frustrante que el Centro de Tesis Electrónico de OhioLINK se conectara bien con otros repositorios y su contenido quedara expuesto a Google y otros motores de búsqueda, pero no hubiera una manera de conectarlo a nuestros propios sistemas de catalogación con facilidad. La única posibilidad era por medio del cliente Connexion del OCLC, que incorpora un proveedor de servicio básico para recuperar registros de repositorios OAI-PMH, pero esto requiere una intervención manual y produce un resultado poco aceptable.

Lo que se hizo fue proponer el desarrollo de un software para crear un registro MARC (legible por máquina) provisional automáticamente en el momento de publicar cada tesis e insertarlo en el catálogo sin intervención humana. La atención de los catalogadores sería mínima, vendría después y se limitaría a revisar los registros y a añadir aquello que no era posible añadir con el proceso automatizado, como la clasificación, la asignación de encabezamientos de materia y la contribución de los registros al OCLC. Debido al lapso de tiempo que transcurriría entre la inserción del registro al catálogo y su revisión final, habría necesidad de contar con una manera de identificar los registros provisionales para que no se olvidaran estos toques finales. Todo el proceso tendría que ser fácil y sencillo y aprovechar lo más posible las capacidades de transferencia de datos ya existente entre los varios sistemas.

Sigamos el flujo de trabajo de una tesis desde la entrega que hace el estudiante mediante una página web. Al terminar la entrega, un correo automático emitido por el Centro de Tesis llega al coordinador de tesis en la biblioteca, quien reenvía el mensaje al decanato correspondiente para ser revisado y aprobado. Al aprobarse la tesis, otro mensaje les llega al coordinador y a los catalogadores. La idea original era que los catalogadores entraran al Centro de Tesis a examinar cada documento y que lo cataloguen como de costumbre. Según la nueva idea, el correo iniciaría un proceso automático.

El primer paso importante para planificar el nuevo proceso implicaría analizar los metadatos disponibles y convertirlos en un formato MARC compatible con el catálogo y con el OCLC. El proveedor de datos

del Centro de Tesis ofrecía dos opciones, el Núcleo de Dublín (exigido por el PMH) y otro esquema, basado en el Núcleo de Dublín pero diseñado para las tesis. Este esquema se llama en inglés ETD (Metadata Schema, Esquema de Metadatos para las Tesis y Disertaciones Electrónicas)[4] también conocido por sus siglas, ETD-MS.

Núcleo de Dublín Core	ETD Metadata (Normas de Metadatos para las tesis y Disertaciones electrónicas) Standard	HTML (Lenguaje de Marcado para Hipertexto)	MARC (Catálogo legible por máquina)
Título	Título	Título	245
Creador	Creador	Creador	100
Descripción	Descripción	Resumen	520
Materia	Materia	Encab. de materia	650
Editor	Editor		
Fecha	Fecha	Fecha	008, 260$c, 502
	Asistente (Asesor)	Asesor	500
	Tipo		
Tipo			
	Formato		538
Idioma	Idioma		008
	Nombre del título	Título	502
	Area de estudio	Título	
	Institución que expide el titulo	Título	502
Identificador	Identificador		035, 856
Derechos (2)	Derechos (2)		
		Páginas	
		Palabras claves	
			256 (tamaño)

4 Networked Digital Library of Theses and Dissertations. ETD-MS: an Interoperability Metadata Standard for Electronic Theses and Dissertations, version 1.00, revision 2, (Biblioteca Digital Electrónica de Tesis y Disertaciones. EDT-MS: Una norma para operar con metadatos para tesis y disertaciones electrónicas, versión 1:00, revisión2)
http://www.ndltd.org/standards/metadata/etd-ms-v1.00-rev2.html (consultado el 29 de septiembre, 2008).

Además, el Centro de Tesis presentaba dos visualizaciones de los metadatos, una página de HTML y una página de texto sencillo que representaba el formato MARC. Como visualizaciones destinadas a la visión humana no eran muy adecuadas para hacer una manipulación programática. Sin embargo vale la pena comparar estas visualizaciones con los esquemas formales de metadatos. Se nota que el esquema ETD-MS tiene un mayor número de elementos y por ser éstos más completos fue adoptado como base de la conversión a MARC.

Una vez elegido el esquema preferido se diseñó un programa denominado el robot catalogador, que serviría como proveedor de servicio con ayuda del OAI-PMH el cual llevaría a cabo las siguientes funciones:

1. Esperar el aviso de publicación.
2. Analizar el correo y extraer el identificador del documento.
3. Recuperar los metadatos del documento.
4. Analizar los metadatos y extraer los datos necesarios.
5. Construir un registro MARC.
6. Enviar el registro al catálogo.
7. Avisar a los catalogadores por correo.

El programa fue puesto en función en junio de 2006 y en general ha funcionado bien. [5]. El mayor inconveniente ha sido el proceso manual de pasar el registro MARC del catálogo al OCLC para completar los últimos pasos de catalogación manual. Otra preocupación es la confiabilidad del correo electrónico que a veces sufre demoras o fallas de servicio.

A partir de septiembre de 2006 el servicio fue dado a conocer en varias reuniones y se inició un diálogo entre los catalogadores del consorcio. Otras dos instituciones habían iniciado procesos parcialmente automatizados y en junio de 2007 se formó un subcomité para considerar todas estas técnicas y crear una encuesta de las institucio-

5 McCutcheon, Sevim, Michael Kreyche, Margaret Beecher Maurer, Joshua Nickerson. Morphing metadata: maximizing access to electronic theses and dissertations (Dándole forma a los metadatos: maximizar el acceso a las tesis y disertaciones electrónicas), Library Hi Tech, vol. 26, no. 1, p. 41-57 (2008).

nes OhioLINK para averiguar si existía interés en una herramienta automatizada de utilidad general para todo el consorcio.

La encuesta tuvo cuatro secciones: Políticas de Catalogación, Contenido de los Registros, Formato de los Registros, y Método de Distribución. De las 12 instituciones que respondieron, 10 tenían la política de catalogar las tesis impresas y sólo 4 la de catalogar las electrónicas. Pero todas las demás 8 incluso las que no habían catalogado las tesis anteriormente, expresaron la intención de iniciar la catalogación de las tesis electrónicas. Las preguntas sobre el contenido de los registros revelaron interés en registrar la fecha de nacimiento del autor para apoyar la creación de registros de autoridades. Las opiniones estuvieron divididas en cuanto a la inclusión del texto estándar y que éste pudiera ser modificado o eliminado según la necesidad de cada caso. En cuanto al formato preferido de los registros, casi todas las respuestas favorecieron el MARC tradicional, hecho no muy sorprendente dado que sólo los sistemas locales y el OCLC apoyan este formato. Los métodos de distribución contemplados fueron OAI-PMH, descarga por enlace web, y la inserción directa que utiliza la Kent State University. Las respuestas no evidenciaron un claro consenso y algunos comentarios señalaron la falta de eficacia en descargar registros uno por uno.

Tomando en cuenta las respuestas, el subcomité formuló cinco recomendaciones principales, aprobadas por el comité plenario en diciembre de 2007:

1. Desarrollar una salida MARCXML.
2. Incorporar enlaces de descarga de registros MARCXML y MARC tradicional (derivado del MARCXML) uno por uno y en lotes conformados de registros previamente marcados.
3. Agregar la salida MARCXML a las opciones DC y ETD-MS del proveedor OAI-PMH.
4. Animar a los catalogadores a utilizar los macros del cliente Connexion of OCLC para agilizar su trabajo.

5. Desarrollar un portal directo si se detecta suficiente interés después de obtener experiencia con los enlaces de descarga.

De estas recomendaciones sólo la primera y la tercera fueron adoptadas. Es decir, el Centro de Tesis ofrece ya suministro de registros MARCXML, disponibles por medio de OAI-PMH, pero la compatibilidad con los catálogos todavía no existe. El rechazo de este punto clave fue causa de cierta frustración y rencor entre los catalogadores.

Resumimos aquí los contrastes culturales antes señalados, entre la cultura de los metadatos y la de la catalogación para ver si son aplicables en esta situación y nos ayudan a entenderla mejor.

Nuevos esquemas vs. el formato MARC

Éste es el contraste que define el conflicto. La cuestión era si el nuevo sistema con sus formatos XML se adaptaría a las necesidades del antiguo, que usa el MARC tradicional. A unos, la concesión de ofrecer el MARCXML sin una conversión a la forma tradicional les pareció una decisión dogmática. Contrariamente esto se hubiera podido justificar con un análisis de recursos laborales o una falta de experiencia con ese tipo de conversión, pero desgraciadamente, la lógica de la decisión no fue comunicada.

Repositorios vs. catálogos o sistemas integrados

Ésta es una división clara que describe la situación, especialmente en relación con los catalogadores, quienes en su papel tradicional no han tenido la oportunidad de ganar experiencia directa con los repositorios.

Tecnólogos vs. bibliotecólogos

Esta distinción también es importante para entender el conflicto. La cultura de metadatos ha surgido a la par que el desarrollo de software, como sucedió en el caso del Centro de Tesis Electrónicas de OhioLINK, ante la falta de productos comerciales. Mientras tanto la catalogación generalmente utiliza productos comerciales. Dicho de otro modo, los catalogadores son consumidores de tecnología y el

personal involucrado con los metadatos tiende a ser creador de la tecnología. Además, el desarrollo del software tuvo lugar en las oficinas del consorcio, un lugar aislado del ambiente bibliotecario.

Recursos inéditos y únicos vs. libros y otros materiales publicados

Esta distinción no cobra mucha fuerza en el caso porqué las tesis no corresponden definitivamente a una categoría u otra. Tradicionalmente se podrían considerar recursos únicos, pero el formato electrónico posibilita su duplicación, y la filosofía de acceso libre la favorece. De igual manera las tesis tradicionales se veían como una clase aparte entre los manuscritos y las publicaciones, pero ahora es más fácil considerar publicadas las tesis electrónicas por su presencia en la web. Puede ser que esta misma naturaleza ambivalente de las tesis haya ocasionado el conflicto cultural, porque las dos comunidades las reivindicaron como territorio propio.

Descripción bibliográfica sencilla y breve vs. completa y precisa

Esta generalización tiene poca relevancia en el caso de estas tesis. Debido a la uniformidad de los documentos, muchos de los datos que deben aparecer en un registro MARC y no existen en los metadatos ETD-MS pueden ser introducidos como valores predeterminados o texto estándar, por ejemplo los campos de longitud fija y algunas de las notas. De hecho los registros creados por el programa generalmente son más largos que los creados manualmente porque incluyen un resumen extenso, generalmente de unas 300 palabras.

Especialistas de la temática vs. especialistas de catalogación

Como se mencionó antes, el autor es la máxima autoridad en cuanto al tema de su tesis, pero como actor transitorio en el proceso de catalogación, no entra en la dinámica del conflicto del presente caso.

OhioLINK ETD Center MARC Record Interface

Select a record set:

- University of Akron ETDs
- Antioch University ETDs
- Ashland University ETDs
- Bowling Green State University ETDs
- Case Western Reserve University ETDs
- Cedarville University ETDs
- Cleveland State University ETDs
- University of Dayton ETDs
- Kent State University ETDs
- Marietta College ETDs
- University of Toledo Health Science Campus ETDs
- Miami University ETDs
- Mount Vernon Nazarene University ETDs
- Ohio Dominican University ETDs
- Ohio University ETDs
- Ohio State University ETDs
- University of Toledo ETDs
- University of Cincinnati ETDs
- Ursuline College ETDs
- Wittenberg University ETDs
- Wright State University ETDs
- Youngstown State University ETDs

or retrieve a record by indentifier: Submit Query

Limit records by creation/modification date:
from 2008 9 28 through 2008 9 28
(Dates should be inclusive but there have been anomalies!)

- (•) Include basic bibliographic data and embargo status
- ( ) Show record numbers and dates only (quicker response)

Submit Query

Repository details

Repository name: OhioLINK Electronic Thesis and Dissertation Center
Base URL: http://www.ohiolink.edu/etd/oai.php
Protocol version: 2.0
Earliest date stamp: 1995-01-01
Record deletion support: no
Granularity of dates: YYYY-MM-DD
Administrative email(s): etd-admin@ohiolink.edu
Compression:
Metadata prefixes supported: oai_dc,oai_etdms,marc21

Nuevo Proveedor de Servicio
(Descarga de Registros MARC)

En suma, tres factores parecen haber contribuido al conflicto: los diferentes formatos, MARC y XML utilizados por los dos grupos; los diferentes tipos de sistemas que tienden a aislar sus usuarios del universo, siempre en aumento, de la bibliotecología; y las distintas relaciones con la tecnología, que tienen el creador y el consumidor.

Como cualquier conflicto cultural, una resolución adecuada depende de la paciencia, la comprensión, y la buena comunicación. Para seguir avanzando hacia una integración de la catalogación de tesis con los metadatos del repositorio, la Kent State ofreció desarrollar un nuevo programa para manejar la descarga y conversión de dichos registros a los catálogos locales. La página principal del prototipo, reproducida aquí, sirve para demostrar las funciones del PMH (Protocolo para cosechar metadatos). En la parte superior aparece el listado de colecciones que corresponde a las instituciones contribuyentes, del cual se selecciona una. Directamente debajo están los límites de los registros por fecha y la opción de recuperar registros breves o completos. En la parte inferior aparece la información que identifica el servidor y sus características.

Este nuevo proveedor de servicio estará al servicio de todas las instituciones de OhioLINK y de cualquier biblioteca que quisiera conseguir registros bibliográficos del Centro de Tesis.