



# El proceso técnico en la biblioteca electrónica

Catalogación y Metadatos

Teresa Malo de Molina

## La colección se va complicando...

### ■ Lo de siempre:

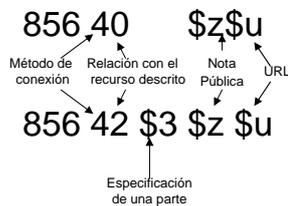
- Libros
- Revistas en papel o Cd-Rom
- Otros materiales no librarios:
  - Partituras
  - Videos
  - Discos
  - ...
- Bases de Datos

### ■ Lo nuevo:

- Revistas electrónicas
- La digitalización de contenidos:
  - Documentos de trabajo
  - Exámenes
  - Tesis
  - ...
- Páginas de contenido:
  - Guías de recursos
  - Artículos y citas concretos
  - ...

## Primera solución: Integrar en el catálogo

- Aprovechamiento del know-how:
  - Lenguaje conocido
  - Fácil inclusión en las tareas diarias
  - Importancia del catálogo como instrumento básico de información
- Posibilidades que ofrece la etiqueta 856



## El cambio es radical

- De objetos físicos a objetos lógicos
- Del acceso físico y único al acceso lógico y múltiple
- Nace la meta-información: un objeto de información en formato digital permite el uso de tecnología también digital para extraer información del propio recurso

## ¿Qué es lo que pasa?

---

### Objetivo

Encontrar información relevante en Internet

### Problema

Los métodos automáticos para identificar recursos en la red (robots de búsqueda y metabuscadores) recuperan gran cantidad de información pero con poca precisión

### Causa

Los documentos de la red carecen de datos suficientes de descripción

## Primer paso para los Metadatos

---

- 1995: Primer Seminario del Dublin Core Metadata
  - Objetivo: llegar a un consenso internacional para diseñar un código de descripción de recursos
  - Intervienen: bibliotecarios, especialistas en lenguajes de marcas, expertos en análisis documental e investigadores de bibliotecas digitales
- Desde entonces, muchas otras asociaciones e instituciones: Library of Congress, OCLC, American Library Association, National Science Foundation, etc.

## ¿Qué son los metadatos?

---

- Los metadatos son datos sobre datos
- Los metadatos son los datos que permiten identificar, describir, analizar y localizar los recursos de Internet
- En el contexto de las bibliotecas son información que:
  - Proporciona – normalmente de forma breve – una caracterización de un recurso de información específico dentro de las colecciones de la biblioteca
  - Se almacena en los catálogos
  - Se usa principalmente para ayudar al usuario a acceder a los recursos de información de su interés
- *Información legible por ordenador sobre recursos web* (Tim Berners-Lee)

## Beneficios de los Metadatos

---

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

## Los metadatos se usan en tres formas distintas...

---

1. Pueden ir acompañando al propio documentos o recurso (en la cabecera de un documento HTML o cabeceras SGML en general)
2. Pueden formar un fichero separado de meta-información para describir recursos distintos a los de HTML (sonido, imagen, programas de ordenador, etc.)
3. Se puede crear con ellos una base de datos central o distribuida con punteros a los recursos que describen

## Evolución de los metadatos

---

- De formatos de estructura muy simple a formatos más complejos
- Desde estándares emergentes propietarios a estándares internacionales
- Los metadatos que se crean se pueden compartir por otros

## Tipos de información que contienen los metadatos

---

1. Identificación y recuperación
2. Condiciones de acceso y requisitos de uso
3. Aspectos estructurales
4. Aspectos contextuales
5. Aspectos de contenido
6. Uso de ese recurso (historia del recurso)

(Bearnan y Sochats)

## ¿Cómo pueden crearse los metadatos?

---

1. Por el propio autor/creador del recurso con la ayuda de herramientas – cada vez más numerosas – creadas “ex profeso” para la generación automática de metadatos
2. Por la organización que gestiona los recursos de información
3. Por un servicio de información, es decir, creadores externos a través de medios humanos (bibliotecarios, por ejemplo) o automáticos (SOIF – Summary Object Interchange Format – en Harvest)

## Características ideales del formato de metadatos

---

- Que sea fácil de crear y de mantener
- Que utilice una semántica que pueda entenderse comúnmente
- Que pueda crearse de forma automática
- Que describa la forma, el contenido y la localización de la información
- Que su estructura permita contenerlos en otros objetos
- Que se puedan usar para construir múltiples índices (www, texto, etc.)
- Que se pueda interoperar en los sistemas de indización que existen
- Que pueda ampliarse según las necesidades

## SGML

---

- Todos los conjuntos de metadatos existentes siguen la norma SGML (Standard Generalized Markup Language = Lenguaje de marcas estándar generalizado)
- Estándar internacional desde 1986 (ISO 8879)

## SGML es un metalenguaje

---

- No es un único lenguaje sino una norma amplia para construir lenguajes de marca
- Proporciona una sintaxis para definir y expresar la estructura lógica de los documentos y las convenciones para nombrar los componentes o elementos de los documentos

## SGML es un metalenguaje

---

- Es un conjunto de reglas formales para definir lenguajes de marcas específicos para tipos específicos de documentos: DTD (Definición del Tipo de Documento):
  - HTML
  - Asociación de Editores Americanos + OCLC:
    - Libros
    - Revistas
    - Artículos
  - Larson (Berkeley) DTD para el USMARC

## MARC vs. SGML

---

- Problemas del formato MARC
  - Está estrictamente controlado, cualquier cambio o adición tarda años en realizarse
  - Es laborioso, lento y costoso de realizar
  - Aunque está compuesto por campos de longitud variable, está limitado a una longitud máxima de 100.000 caracteres
  - Se adapta muy mal a la información estructurada jerárquicamente

## MARC vs. SGML

---

- Ventajas del SGML
  - Puede tratar información jerárquicamente interrelacionada en tanto niveles como se necesite
  - No tiene limitación en el tamaño de los registros
  - Es un estándar internacional adoptado por un número creciente de instituciones (más allá de las bibliotecas)
  - Permite una flexibilidad máxima en el uso del texto
  - Los registros pueden estar interrelacionados en distintos ficheros

## Algunos proyectos de Metadatos

---

- DC (Dublin Core Metadata)  
<http://dublincore.org>
- RDF (Resource Description Framework)  
<http://www.w3.org/RDF/>
- TEI (Text Encoded Initiative)  
<http://www.tei-c.org>
- URC (Uniform Resource Characteristic/Citations)
- MARC DTD (Machine Readable Cataloging Document Type Definition)
- EAD (Encoded Archival Description)  
<http://www.loc.gov/ead/>
- IMS (Instructional Management System)  
<http://www.imsproject.org>

## DUBLIN CORE METADATA

---

- 1995: Dublin Metadata Workshop
- Abril 1996: Warwick Metada Workshop
- Septiembre 1996: Image Metada Workshop
- 1997: Camberra (Australia)

## DUBLIN CORE METADATA

### ■ Resultados:

- Una sintaxis concreta para el Dublin Core, expresada como DTD  $\implies$  se mapeó a etiquetas HTML para que los metadatos se pudieran insertar en documentos web
- Warwick Framework, arquitectura que permite reunir los diferentes paquetes de metadatos, pudiendo acceder a ellos y mantenerlos
- Una guía para el usuario con nivel básico y complejo
- Extensión del Dublin Core para recursos visuales e imágenes digitales

## LOS 15 ELEMENTOS DEL DUBLIN CORE

Metadatos relativos al contenido	Metadatos relativos a la propiedad intelectual	Metadatos en relación con la autoridad documental
<ul style="list-style-type: none"><li>•Título</li><li>•Materia y palabra clave</li><li>•Descripción</li><li>•Fuente</li><li>•Idioma</li><li>•Relaciones: enlaces a otros recursos</li><li>•Cobertura</li></ul>	<ul style="list-style-type: none"><li>•Autor</li><li>•Editor</li><li>•Autor secundario</li><li>•Gestión de derechos</li></ul>	<ul style="list-style-type: none"><li>•Fecha</li><li>•Tipo de recurso</li><li>•Formato</li><li>•Identificador del recurso</li></ul>

## Proyectos existentes que aplican DC

---

- La Biblioteca Nacional de Australia
- Nordic Metadata Project
- Distributed Systems Technology Centre (Australia)
- Library of Congress : Programa Nacional de la Biblioteca Digital
- Desarrollo del estándar Z39.50

## Ejemplo

---

Subject:  
  scheme=keywords:Electronically mediated environments  
    Cyberspace  
    Urbanism  
    Architecture  
  scheme=LCSH:Computer networks  
    Information technology  
    Virtual reality  
    Computers and civilization  
Title:City of Bits: Space, Place, and the Infobahn  
Author:Mitchell, William J.  
Publisher: MIT Press  
Other Agents:  
  otherAgent role=WWW team member: Stevenson, Daniel C.  
  otherAgent role=WWW team member: Ehling, Teresa  
Date: 1995  
Identifiers:  
  scheme=ISBN: 0262133091  
  scheme=URL: [http://www.mitpress.mit.edu:80/City\\_of\\_Bits/WWW/Preamble.html](http://www.mitpress.mit.edu:80/City_of_Bits/WWW/Preamble.html)  
Object type:book  
Form:Text/HTML, Video/(MPEG, Quicktime). Image/GIF  
Language:English  
Source:type=print ed.: City of Bits: Sapce, Place, and Infobahn

## RESOURCE DESCRIPTION FRAMEWORK

---

- Se inicia a mediados de 1997 por el W3C (World Wide Web Consortium)
- No es sólo un formato de metadatos, sino un marco genérico de descripción de recursos
- Su objetivo principal es proporcionar interoperabilidad entre aplicaciones que intercambian información entendible por ordenador en la web, es decir, crear un lenguaje (modelo y sintaxis) para el intercambio de descripciones de recursos web

## RESOURCE DESCRIPTION FRAMEWORK

---

- Puede utilizarse en gran variedad de áreas:
  - En un motor de búsqueda para incrementar y mejorar la recuperación
  - En catalogación para describir el contenido y las relaciones de contenido disponibles en una sede web, en una página o en una biblioteca digital
  - Para describir los derechos de propiedad de las páginas web
  - Para expresar los permisos de accesibilidad de los usuarios y de las sedes web

## RDF vs. DC

---

- La originalidad y mayor capacidad de RDF reside en:
  - Permite especificaciones semánticas y una sintaxis para múltiples formatos de metadatos (DC incluido)
  - Se integra en la estructura XML (Extensible Markup Language) que enriquece sus posibilidades de uso – el W3C está trabajando con los diseñadores de DC para implantarlo en un entorno XML
  - Su definición XML cuenta con el apoyo de Netscape y Microsoft
- Estas características lo convierten en el estándar más prometedor para búsquedas y recuperación de la información. De hecho NISO e ISO lo apoyan

## TEXT ENCODING INICIATIVE

---

- Participan tres instituciones:
  - Association for Computers Humanities
  - Association for Computational Linguistics
  - Association for Literary and Linguistic Computing
- Inicialmente pretende desarrollar un DTD para los textos que se utilizan en lingüística, literatura y estudios históricos, pero actualmente pretende buscar el sistema que permita a los investigadores de todas las disciplinas intercambiar y reutilizar recursos, independientemente del software y hardware que utilicen y sin tener en cuenta dónde están localizados

## CARACTERISTICAS DEL TEI

---

- Todos los textos TEI deben ir precedidos de una cabecera (TEI Header) que describa el texto
- Un comité específico – bibliotecarios y archiveros de Europa y USA – se encarga de diseñar las especificaciones de esta cabecera
- Su propósito es usar la cabecera como medio de control bibliográfico

## CARACTERISTICAS DEL TEI

---

- La cabecera puede almacenarse como parte separada del documento al que se refiere o ir unida intrínsecamente a él
- Se puede utilizar además para describir otros recursos en la red cuando sea necesario
- Su gran flexibilidad permite adaptarlo fácilmente a cualquier usuario, ya que permite un mayor o menor nivel de detalle en la descripción

## Tipos de información de la cabecera TEI

---

1. Descripción del fichero: características bibliográficas del documento
2. Descripción codificada: que muestra cómo se codificó el texto
3. Información descriptiva: información adicional contextual y no bibliográfica: idioma, colaboradores, materias, descriptores, etc.
4. Descripción histórica sobre los cambios que ha sufrido el texto electrónico: detalle sobre recursos, adiciones al texto, etc.

Las descripciones que se proporcionan para cumplimentar los elementos más estructurados siguen las AACR2 y las ISBD. Los elementos no estructurados contienen texto libre

## Implementaciones del TEI

---

- Oxford Text Archive  
<http://sabel.ox.ac.uk/ota>
- Electronic Text Centre (Universidad de Virginia)  
<http://www.lib.virginia.edu/etext/ETC.html>
- Conversión de cabeceras TEI a USMARC (Biblioteca de la Universidad de Virginia)

## UNIFORM RESOURCE CHARACTERISTIC/CITATIONS

- Promovido por el Grupo de Trabajo de Internet Task Force's Uniform Resources Identifiers
- Se ha utilizado generalmente para identificar:
  - Información catalográfica de los recursos en línea
  - Una normalización de metadatos pero no necesariamente con propósitos catalográficos
  - Información que puede utilizarse en el proceso de asignar Nombre de Recurso Uniforme (URN: Uniform Resource Name) → propuesta para asignar localizaciones únicas independientes a los recursos en la red (parecido al ISBN)
- Se ha concebido como un servicio general al que las personas autorizadas pueden acceder para modificar la URL asociada al URN

## Elementos del URC

- Puede incluir metadatos que pueden ayudar a describir el recurso y recuperarlo
- Puede incluir además otro tipo de elementos como firma electrónica e información revisada para averiguar la veracidad del recurso
- Puede incluir un elemento de acceso para restricciones de uso
- Puede incluir un elemento que indique la versión e historia del documento

## MARC DTD PROJECT

---

- Proyecto de la Library of Congress
- Objetivo: conseguir un DTD que permita la conversión de los registros MARC a una estructura SGML y viceversa, sin pérdida de datos
- Incluye dos partes:
  - Desarrollo del DTD correspondiente a los tres formatos USMARC: de datos bibliográficos, de datos de autoridad y de fondos y localizaciones
  - Desarrollo de un software que permita la conversión automática entre los dos formatos

## MARC DTD PROJECT

---

- Se inicia en 1995
- Versión alpha del DTD disponible en mayo de 1996
- Versión definitiva del software disponible en enero de 1998
- La Library of Congress permite descargarlo de forma gratuita:  
<http://www.loc.gov/marc/marcdtd/marcdtdbeta.html>

## CONCLUSIONES

---

- La tendencia a la utilización de metadatos es imparable
- El OPAC puede servir de pasarela para acceder a una base de datos de metadatos en Internet
- Se tendrían que establecer perfiles Z39.50 para cada uno de los formatos de metadatos
- El papel del bibliotecario es contribuir a la investigación de sistemas y estándares de futuro
- No caigamos en la tentación de buscar nuevos nichos de trabajo ahora que el Z39.50 nos ahorra la catalogación original