Digital Libraries: Catalogues

RAFFAELLA BERNARDI

UNIVERSITÀ DI TRENTO

P.ZZA VENEZIA, ROOM: 2.05, E-MAIL: BERNARDI@DISI.UNITN.IT

Contents

1 OPAC 5
1.1 Catalogue Card at my time 6
1.2 Catalogue Card at your time 8
1.3 WorldCat 9
2 Metadata 10
2.1 Where metadata come from and where they are 11
2.2 Different types of metadata 12
2.3 Metadata Format: MARC 13
2.3.1 MARC: an example 14
2.3.2 MARC: an example (legenda) 15
2.3.3 Actual Marc record 16
2.3.4 Tags 17
2.3.5 Evolution of MARC format 18
2.4 Set of metadata elements: Dublin Core 19
2.4.1 DC simple: 15 elements 20
2.4.2 MARC vs. DC 21
2.4.3 Metadata for citation: Bibtex 22

2.4.4 Metadata for Citation: RIS 23
2.5 Metadata syntax: RDF 24
3 What and How to write in the metadata values 25
3.1 Cataloging Rules 26
3.2 Need of Authority Control 27
3.3 Authority Files 28
4 Subject Headings 29
4.1 Controlled Vocabulary for SHs 30
4.2 Library of Congress Subject Headings 31
4.3 Subject Heading Systems: Lists 32
4.4 Example: SOG vs. LCSH 34
5 Classification System 35
5.1 Library of Congress (up to 1892) 36
5.2 Library of Congress (today) 37
5.3 Dewey Decimal System 38
5.4 Classification Systems: List 39
5.5 New trends: Google categories 40
5.6 Example: Record 41
6 Conclusions 42

1. OPAC

OPAC (Online Public Access Catalog) is an online data base (viz. a database accessible from a network) of material held in a Library (or group of libraries).

A *library catalog* is a register of all bibliographic items found in a library. In 1876 Charles Ammi Cutter defined the objectives of a bibliographic system to be:

- to enable a person to find a book of which either the author, the title, the subject or the category is known. [*Identifying* objective]
- to show what the library has by a given author, subject, kind of literature [*Collocating* objective]
- to assist in the choice of a book [*Evaluating* objective]

Library Catalogs originated as manuscript *list* arranged alphabetically by author. The first *card catalogue* appeared in the nineteenth century, enabling much more flexibility, and towards the end of the twentieth century the OPAC was developed.

1.1. Catalogue Card at my time

1.2. Catalogue Card at your time

Name:	Arnosky, Jim.
Title proper:	Raccoons and ripe corn.
Statement of responsibility:	Jim Arnosky
Edition statement:	1st ed.
Place of publication:	New York
Name of publisher:	Lothrop, Lee &Shepard Books
Date of publication:	c1987
Pagination:	25 p.
Illustrative matter:	col. ill.
Size:	26 cm
Summary:	Hungry raccoons feast at night in a field of ripe corn
Topical subject:	Raccoons
Local call number:	599.74 ARN
Local barcode number:	8009
Local price:	$15.00

1.3. WorldCat

WorldCat http://www.oclc.org/worldcat/ is presently the biggest OPAC:

- about 10 thousand libraries from more than 90 countries
- more than 90 million records
- 1200 million physical and digital assets 360 languages

run by Online Computer Library Center (OCLC, US organization) that helps libraries locate, acquire, catalog, and lend library materials.

The *sharing of metadata* is made possible by the use of standardized records.

2. Metadata

Metadata are “data about data”. It’s structured information about a particular information resource. When an information is “structured” it can be manipulated without understanding its content. Important questions are:

- Where does the metadata come from (automatically extracted vs. manually assigned vs. imported)?
- How will the metadata affect the document display, browsing, searching, and maintenance of the digital library?
- Does it need harmonization (e.g. different versions of people’s names.)
- Is the metadata private to the library or can it be shared with others?

2.1. Where metadata come from and where they are

The principles of metadata for a Library and a Digital Library are the same.

- Where they come from:
  - Human-assigned metadata
  - If the document “was born digital”, metadata may have been embedded within the file at the moment of its creation.
  - Automatically assigned metadata: a program process the digital document and output a value for the metadata element.
- Where they are:
  - many file formats have embedded metadata
  - stored separately in a library catalog

2.2. Different types of metadata

In a Library there are different types of metadata:

- Administrative metadata: managing resources.
- Descriptive metadata: describing resources.
- Technical metadata: low-level system information (e.g. data-format, data compression used.)
- Usage metadata: related to system use (e.g. tracking user behavior)

The move from card catalogs to computer-based records asked for tools to manage metadata.

2.3. Metadata Format: MARC

MARC (Machine readable cataloging) format was developed in the 1960s by Henriette Avram at the Library of Congress.

It was originally used to automate the creation of physical catalog cards.

It provides the protocol by which computers *exchange*, use, and interpret bibliographic information. Its data elements make up the foundation of most library catalogs used today.

It contains several hundred elements. For example, 6XX fields are for subject headings; 600, if it's a person; 610, if it's a corporation; 651, if it is a place, ...

Many "national" versions (UKMARC, CANMARC, AUSMARC, DANMARC, ANNA-MARC, INTERMARC, etc) and UNIMARC (Universal MARC) as standard format for exchange of information.

Recall Disks were small, and their costs where high. There was the need of using codes of small bytes. The same for processing the documents (numbers instead of words were easier to be processed). The same for sharing the data, etc.

2.3.1. MARC: an example

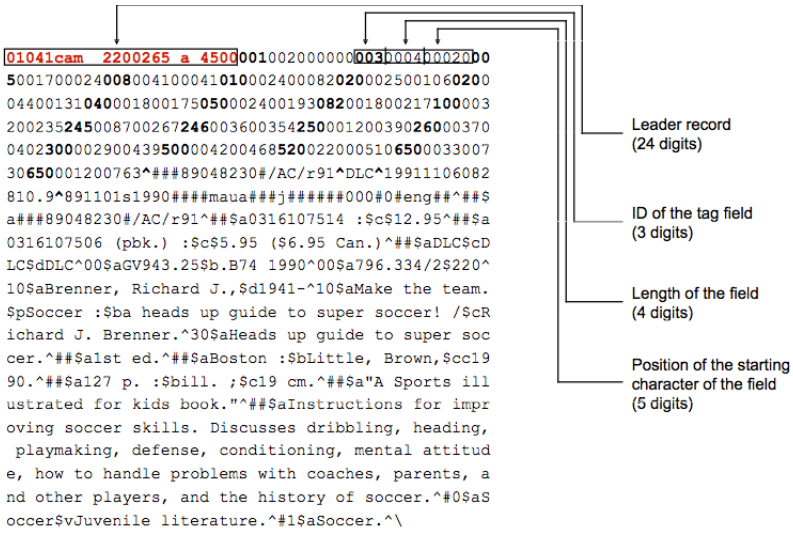
"SIGNPOSTS"			DATA
100	1#	$a	Arnosky, Jim.
245	10	$a $c	Raccoons and ripe corn / Jim Arnosky.
250	##	$a	1st ed.
260	##	$a $b $c	New York : Lothrop, Lee & Shepard Books, c1987.
300	##	$a $b $c	25 p. : col. ill. ; 26 cm.
520	##	$a	Hungry raccoons feast at night in a field of ripe corn.
650	#1	$a	Raccoons.
900	##	$a	599.74 ARN
901	##	$a	8009
903	##	$a	$15.00

2.3.2. MARC: an example (legenda)

```
300 ## $a 675 p. : $b ill. ; $c 24 cm.
```

- Tag 300 means a book's physical description
- `##` means no indicators
- Subfield `$a` indicates the extent (number of pages)
- Subfield `$b` indicates other physical details (illustration information)
- Subfield `$c` indicates dimensions (centimeters)
- the character `#` is the place-holder for indicators the character `$` indicates the beginning of a sub-field

2.3.3. Actual Marc record



2.3.4. Tags

Tag	Length	Starts at	Tag	Length	Starts at
001	0020	00000	100	0032	00235
003	0004	00020	245	0087	00267
005	0017	00024	246	0036	00354
008	0041	00041	250	0012	00390
010	0024	00082	260	0037	00402
020	0025	00106	300	0029	00439
020	0044	00131	500	0042	00468
040	0018	00175	520	0220	00510
050	0024	00193	650	0033	00730
082	0018	00217	650	0012	00763

2.3.5. Evolution of MARC format

MARC 21 is a result of the combination of the United States and Canadian MARC formats (USMARC and CAN/MARC). MARC 21 was designed to redefine the original MARC record format for the 21st century and to make it more accessible to the international community.

MARC XML is an XML schema based on the fairly common MARC 21 flavour. It was developed by the US Library of Congress and adopted by it and others as a means of easy sharing of, and networked access to, bibliographic information. Being easy to parse by various systems allows it to be used as an aggregation format, as it is in software packages such as MetaLib. E.g.:

http://opac.bncf.firenze.sbn.it/opac/controller.jsp?action=notizia_viewxml&notizia_idn=umc0295572

2.4. Set of metadata elements: Dublin Core

Dublin Core is a *set of predefined metadata elements* intended for the description of electronic material. Implementations of Dublin Core typically make use of XML and are Resource Description Framework (RDF) based.

The “Dublin” in the name refers to Dublin, Ohio, U.S., where the work originated from an invitational workshop hosted by OCLC.

The “Core” refers to the fact that the metadata element set is a basic but expandable ”core” list.

- *Simple Dublin Core* comprises *fifteen elements* (Title, Creator, Subject, Description, Publisher . . .); each Dublin Core element is *optional* and may be *repeated*.
- *Qualified Dublin Core* includes *additional elements* (Audience, Provenance and RightsHolder, etc.), as well as a group of element refinements (also called *qualifiers*) that refine (narrower) the semantics of the elements in ways that may be useful in resource discovery. They can be ignored if not understood by a machine.

2.4.1. DC simple: 15 elements

Content	Intellectual Property	Instantiation
Title	Creator	Date
Subject	Contributor	Format
Description	Publisher	Identifier
Type	Rights	Language
Source		
Relation		
Coverage		

2.4.2. MARC vs. DC

MARC is a comprehensive, well-developed, carefully controlled scheme intended to be generated by professional catalogers for use in libraries.

Dublin Core is an intentionally minimalist standard intended to be applied to a wide range of digital library materials by people who are not trained in library cataloging.

These two schemes are of interest not only for their practical value, but also to highlight diametrically opposed underlying philosophies.

2.4.3. Metadata for citation: Bibtex BibTex is a package of Tex/LaTex to manage bibliographic data within document. Records begin with `@` symbol followed by a keyword naming the record type

```
@Article{,
  author =   {},
  title =   {},
  journal =   {},
  year =   {},
  OPTkey =   {},
  OPTvolume =   {},
  OPTnumber =   {},
  OPTpages =   {},
  OPTmonth =   {},
  OPTnote =   {},
  OPTannote =   {}
}
```

2.4.4. Metadata for Citation: RIS EndNote is a bibliographic tool. Its format can be converted into HTML, XML etc. and many systems can export its format. For instance, it handles RIS.

RIS is a tagged format for expressing bibliographic citations developed by the Research Information Systems. Eg.:

```
TY  - JOUR
AU  - Shannon,Claude E.
PY  - 1948/07//
TI  - A Mathematical Theory of Communication
JO  - Bell System Technical Journal
SP  - 379
EP  - 423
VL  - 27
ER  -
```

2.5. Metadata syntax: RDF

RDF (Resource Description Framework) is designed to facilitate the interoperability of metadata. It supplies a means for describing a valid system (by describing the syntax to define an RDF schema). The basic construction is a binary relation that connect a subject to and object, e.g. Title(isbn:9780..,"how to build..") and form a statement. The isbn works as the URI (Universal Resource Identifier) of the record.

DC defines the elements to be filled in, RDF provides the syntax to be used to write these elements.

```
<?xml version="1.0"?><rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc= "http://purl.org/dc/elements/1.1/">
<rdf:Description rdf:about="http://www.w3schools.com">
  <dc:description>W3Schools - Free tutorials</dc:description>
  <dc:publisher>Refsnes Data as</dc:publisher>
  <dc:date>2008-09-01</dc:date>
  <dc:type>Web Development</dc:type>
  <dc:format>text/html</dc:format>
 </rdf:Description>
</rdf:RDF>
```

3. What and How to write in the metadata values

The first three elements of DC are: 1. Title, 2. Creator. 3. Subject.

For each of them there are *exact rules* on how their values should be written as well as *Controlled Vocabularies* providing the term themselves.

3.1. Cataloging Rules

Cataloging rules have been defined to allow for consistent cataloging of various library materials.

The most commonly used set of cataloging rules in the English speaking world are the Anglo-American Cataloguing Rules, 2nd Edition revised (AACR2R) Example: Rules to know how to name a local church; choose the name in this order

1. “the person(s), object(s), place(s), or event(s) to which the local church [..] is dedicated or after which it is named.”
2. “A name beginning with a word or phrase descriptive of a type of local church”
3. “A name beginning with the name of the place in which the local church [..] is situated.”

3.2. Need of Authority Control

In some catalogs, person's names are *standardized*, i.e., the name of the person is always (cataloged and) sorted in a *standard form*, even if it appears differently in the library material.

Table 2.1 Spelling variants of the name Muammar Qaddafi.

Qaddafi, Muammar	Muammar al-Qadhafi	Qathafi, Muammar
Gadhafi, Mo ammar	Mu ammar al-Qadhdhafi	Gheddafi, Muammar
Kaddafi, Muammar	Qadafi, Mu ammar	Muammar Gaddafy
Qadhafi, Muammar	El Kazzafi, Moamer	Muammar Ghadafi
El Kadhafi, Moammar	Gaddafi, Moamar	Muammar Ghaddafi
Kadhafi, Moammar	Al Qathafi, Mu ammar	Muammar Al-Kaddafi
Moammar Kadhafi	Al Qathafi, Muammar	Muammar Qathafi
Gadafi, Muammar	Qadhdhafi, Mu ammar	Muammar Gheddafi
Mu ammar al-Qadafi	Kaddafi, Muammar	Khadafy, Moammar
Moamer El Kazzafi	Muammar al-Khaddafi	Qudhafi, Moammar
Moamar al-Gaddafi	Mu amar al-Kad'afi	Qathafi, Mu'Ammar el
Mu ammar Al Qathafi	Kad'afi, Mu amar al-	El Qathafi,Mu'Ammar
Muammar Al Qathafi	Gaddafy, Muammar	Kadaffi, Momar
Mo ammar el-Gadhafi	Gadafi, Muammar	Ed Gaddafi, Moamar
Muammar Kaddafi	Gaddafi, Muammar	Moamar el Gaddafi
Moamar El Kadhafi	Kaddafi, Muamar	

3.3. Authority Files

This standardization is achieved by a process called *authority control* (viz. the practice of creating and maintaining index terms for bibliographic material in a catalog). This is usually done for authors' names and titles.

The most common way of enforcing authority control in a bibliographic catalog is to set up a separate index of authority records, which relates to and governs the headings used in the main catalog. This separate index is often referred to as an *authority file*.

Project on Authority File: VIAF (Virtual International Authority File):
http://viaf.org/

4. Subject Headings

Their definition A subject heading is a term that captures the essence of the *topic* of a document.

Their use They are used as keywords to provide *subject access points* to the bibliographic records contained in Library catalogs.

SHs can consist of a word or a phrase. They are created by analyzing the document either manually or automatically. They can either come from a controlled vocabulary or be freely assigned.

4.1. Controlled Vocabulary for SHs

Authority Files contain the authorized forms of subject headings, their synonyms, and related subject headings.

Subject heading classification is a human and intellectual endeavor, where trained professionals apply topic descriptions to items in their collections.

Multiple SHs may be assigned to a single resource, thereby providing multiple access points for each resource. SHs can be highly specialized and fine-grained.

Some Subject Headings Lists are *thesauri*, and supply information on hypernyms, hyponyms and related terms.

4.2. Library of Congress Subject Headings

Agricultural Machinery

- UF (use for):
  - Agriculture – Equipment and supplies
  - Crops – Machinery
  - Farm Machinery
- BT (broader term)
  - Machinery
- RT (related term)
  - Farm Equipment
  - Farm Mechanization
- SA (see also)
  - subdivision Machinery under names of crops (e.g. Corn – Machinery)
- NT (narrower term)
  - Agricultural implements
  - Agricultural instruments

Agriculture – Equipment and supplies
USE Agriculture Machinery

Crops – Machinery
USE Agriculture Machinery

Farm Machinery
USE Agriculture Machinery

4.3. Subject Heading Systems: Lists

- Library of Congress of Subject Headings (LCSH): http://id.loc.gov/search/
- Soggettario Italiano (SOG): http://thes.bncf.firenze.sbn.it/ricerca.php
- Schlagwortnormdatei (SWD)
  http://www.d-nb.de/standardisierung/normdateien/swd.htm
- Rameau http://rameau.bnf.fr/
- ...

4.4. Example: SOG vs. LCSH

Biblioteche Digitali (SOG)

RDF/XML

Biblioteche digitali Gerarchia Notizie bibliografiche

Categoria/Faccetta: Cose:Strumenti
Nota d'ambito: Sistemi di risorse elettroniche organizzate in spazi virtuali e fisici, in cui interagiscono funzioni di progettazione e sviluppo dei servizi di acquisizione, manutenzione e preservazione di documenti digitalizzati o nati in formato digitale, conservati in appositi depositi digitali

TT	Strumenti
BT	[Strumenti relativi alla biblioteconomia, all'archivistica]
RT	Archivi
	Archivi di dati
	Biblioteche
	Digitalizzazione
	Risorse elettroniche

Fonte: BGC; →Wikipedia; DDC22; →LCSH: Digital libraries
Classificazione Dewey (Ed. 22): 025.04

Agenzia catalografica/Proponente: BNI
Status del record: Termine strutturato
Identificativo: 30683

Digital Libraries (LCSH)

Alternate Labels
- Digital curation
- Digital media collections
- Digital media libraries
- Electronic libraries
- Electronic publication collections
- Electronic publication libraries
- Electronic text collections
- Virtual libraries

Broader Terms
- Libraries

Narrower Terms
- Art--Digital libraries
- Australia--Digital libraries
- Children's digital libraries
- Computer science--Digital libraries
- Education, Higher--Digital libraries
- Education--Digital libraries
- Europe, Eastern--Digital libraries
- Humanities--Digital libraries
- Institutional repositories
- Korea--Digital libraries
- Science--Study and teaching--Digital libraries
- Slavic countries--Digital libraries
- Social sciences--Digital libraries
- Tasmania--Digital libraries
- Z39.50 Profile for Access to Digital Collections

Related Terms
- Information storage and retrieval systems
- Web archives

5. Classification System

Classification System Classification means to bring related items together. Conventional libraries, in order to *stack* books on related subjects together, have used library classification.

In the past the main task of CS has been to bring related items together in a helpful *sequence* from the general to the more specific. This might include shelving materials in CS-order.

For this reason, in many libraries, only *one* CS notation is provided for a resource.

Moreover, the notation is unambiguous and language independent since it consists of *conventional sequences of letters, numbers, and/or punctuation*.

The notations correspond to topics, expressed in the language of the country where the CS originated, and are organized *hierarchically by disciplines*.

5.1. Library of Congress (up to 1892)

- Sacred history
- Ecclesiastical history
- Civil history
- Geography, travels
- Law
- Ethics
- Logic, rethoric, criticism
- Dictionaries, grammars
- Politics
- Trade, commerce
- Military and naval tactics
- Agricolture
- Natural history
- Medicine, surgery, chemistry
- Poetry, drama, fiction
- Arts, sciences, miscellaneous
- Gazettes (newspapers)
- Maps

In 1812 the Library of Congress moved to a “new” classification scheme, with 44 categories, and then, in 1897, to the present scheme

5.2. Library of Congress (today)

A - General Works
B - Philosophy, Psychology, Religion
C - Auxiliary Sciences of History
D - History: General & Outside the Americas
E - History: United States
F - History: United States Local & America
G - Geography, Anthropology, Recreation
H - Social Sciences
J - Political Science
K - Law
L - Education
M - Music
N - Fine Arts
P - Language and Literature
Q - Science
R - Medicine
S - Agriculture
T - Technology
U - Military Science
V - Naval Science
Z - Library Science & Information Resources

5.3. Dewey Decimal System

000 – Computer science, information, and ger
100 – Philosophy and psychology
200 – Religion
300 – Social sciences
400 – Languages
500 – Science and Mathematics
600 – Technology and applied science
700 – Arts and recreation
800 – Literature
900 – History and geography and biography

http://opac.bncf.firenze.sbn.it/opac/controller.jsp?action=dewey_browse

5.4. Classification Systems: List

- Library of Congress (LoC): http://www.loc.gov/catdir/cpso/lcco/
- Dewey Decimal System (DDC) (reference point for international mappings) http://www.oclc.org/dewey/
- Regensburger Verbundklassifikation (RVK) http://rvk.uni-regensburg.de/
- Göttinger Classification system (GOK)
- ...
- See http://renardus.sub.uni-goettingen.de/renap/racr.html for more.

5.5. New trends: Google categories

- **Arts**
  - Movies, Music, Television, ...
- **Business**
  - Industries, Finance, Jobs, ...
- **Computers**
  - Hardware, Internet, Software, ...
- **Games**
  - Board, Roleplaying, Video, ...
- **Health**
  - Alternative, Fitness, Medicine, ...
- **Home**
  - Consumers, Homeowners, Family, etc...
- **Kids and Teens**
  - Computers, Entertainment, School, etc...
- **News**
  - Media, Newspapers, Current Events, ...
- **Recreation**
  - Food, Outdoors, Travel, ...
- **Reference**
  - Education, Libraries, Maps, ...
- **Regional**
  - Asia, Europe, North America, etc...
- **Science**
  - Biology, Psychology, Physics, etc...
- **Shopping**
  - Autos, Clothing, Gifts, ...
- **Society**
  - Issues, People, Religion, ...
- **Sports**
  - Basketball, Football, Soccer, ...

Others: e.g. Yhaoo categories, DBPedia categories..

5.6. Example: Record

Michael Lesk “Understanding Digital Libraries”

- FUB

  SH Elektronische Bibliothek Libraries [SWD] / United States / Special collections / Computer files Digital libraries / United States [LCSH]

  CS AN 73000 [RVK]

- Göettingen University Library

  SH Elektronische Bibliothek Libraries / Special collections / Computer files / United States Digital libraries / United States

  CS (shelf mark) 2005 A 12925.

  Further CS Z692.C65 [LoC] and 025.00285 [DDC]

- CNR Pisa

  SH Information Storage and retrieval, Digital Libraries [ACM Subject Terms]

  CS H.3.7 [ACM classification]

6. Conclusions

We have seen that the catalogue card may:

- be written in different formats (MARC, Dublin Core/Qualified DC)
- contain different subject headings (LCSH, SOG, etc..)
- contain different Classification system (DDC, RVK, etc.)

all these differences call for "interoperability": (a) Formats need to be converted, (b) SH and CS need to be mapped. (a) is done for most formats, (b) has been tackled but it's a hard problem.

- Libraries: Only data available are the metadata – they are pretty rich and reliable.
- Digital Libraries: The metadata could be poor and less reliable but besides the metadata there is free text (abstract or document).