

Introduction to Digital Libraries

hussein suleman
uct cs honours 2003

The Library Perspective

The Metadata Bottleneck

- Historically, libraries have created metadata by hand for each item – manual.
 - Expensive
 - Time-consuming
 - Human-intensive
- Computer scientists argue for automatic metadata extraction.
 - Cheaper and faster
 - Lower quality
 - Based on heuristics instead of human experts

The Search Dilemma

- Searching has traditionally been based on carefully selected keywords.
 - A good keyword
 - does not appear in the title or abstract
 - is specific and descriptive but not obscure
- Modern users expect to use Google-like search engines based on automatic indexing of full-text rather than keywords.
- Have you ever been asked for keywords when writing a document?

Direct or Indirect Retrieval

- Information retrieval in libraries often involves an intermediary (e.g., subject librarian) who performs the search on your behalf. The hypotheses are:-
 - Search engines, like ALEPH, are too complicated and cannot be used effectively by users!
 - Most users cannot construct effective queries to obtain the best possible results.
- Knowing what you know about IR, how does one construct an effective query?

Quality Control

- Librarians obtain material from publishers, who have a network of book/journal editors/reviewers to maintain quality.
- How is quality control enforced online?
- How valid is an online reference compared to a reference to a published document?
- Why publish electronic pre-prints if the quality is lower? How do we maintain a high standard of pre-prints?
- How does an establishment like a library decide what electronic material to vet?

Authority Control

- What if two people share the same name – how do you “search for other books by the same author”?
 - Use dates of birth
 - Use unique/global identity numbers
 - Assign authority numbers
 - Use email addresses?

- How important is this?

Another hussein?

Councillor Hussein Suleman
(Liberal Democrat)



Tel: 0116 252 6045 - Leicester City Council

Surgeries:

Evington Valley Renewal Office, 166 Evington Road (near Co-op), 2nd Friday each month, 6.00pm - 7.00pm.

Medway Junior School, St Stephens Road, 4th Wednesday each month, 6.00pm-7.00pm

[Click here to contact Councillor Hussein Suleman Online](#)

Library Consortia

- ❑ Cataloguing is cooperative – why should two institutions catalogue the same book in a networked environment?
- ❑ UCT is part of CALICO (Cape Library Cooperative), to coordinate and broaden services and reduce costs.
- ❑ OCLC's WorldCat contains 52 million resource descriptions that are developed cooperatively.
 - UCT is a member!
- ❑ SABINET runs SACat – a local equivalent.

Classification Systems

- ❑ Classifications are used to organise collections of information into categories (usually hierarchically).
- ❑ Dewey Decimal System (OCLC)
 - Based on a decimal numbers. e.g., 004.67
 - Computer Science is in 001
(000=Information)? 520 (500=Sciences)? 620
(600=Technology)?
- ❑ Library of Congress Classification
 - Based on alphanumeric/decimal codes e.g., QA76.76
 - Computer Science is in QA76
(QA=Mathematics)

Metadata Standards - MARC

- ❑ MACHine-Readable Cataloguing is an abstract metadata format to describe library items, where each field and subfield is specified by a number.
- ❑ Format used by library catalogue software, such as ALEPH.
- ❑ Various different standards – USMARC, UKMARC, CAN/MARC, etc.
- ❑ Current standard version – MARC21
- ❑ XML encoding used in OAI data providers.

MARC Example 1

The screenshot displays the ALEPH library system interface. At the top, there is a navigation bar with the ALEPH logo and links for Home, Database, Records, Display, and Help. Below this, there are buttons for Record, Edit Record, Create, Renewal, and Edit Log. A message indicates the user is viewing Record 16 out of 48. The main area shows a MARC record in text format, starting with 'FMT BK' and 'LDR 01500a1 0200 1:450'. The record includes various fields such as 001 (00125817), 005 (200510200901C), 008 (00) (0200) (medium), 010 (650477), 020 (65001625), 020 (650511774), 036 (6504108), 036 (650510254), 040 (1), 08200 (650510254), 1001 (650510254), 21510 (650510254), 260 (650510254), 300 (650510254), 440 0 (650510254), 504 (650510254), 505 (650510254), and 650 0 (650510254). The record describes a book titled 'The name and life of the first Pope' by the Vatican Library.

MARC Example 2

```
<marc:record xmlns:marc="http://www.loc.gov/MARC21/slim"
  type="Bibliographic">
  <marc:leader>01476ckm 22003737a 4500</marc:leader>
  <marc:controlfield tag="001"> 2002722378</marc:controlfield>
  <marc:controlfield tag="003">DLC</marc:controlfield>
  <marc:controlfield tag="005">20030220135251.0</marc:controlfield>
  <marc:controlfield tag="007">kj bo </marc:controlfield>
  <marc:controlfield tag="007">cr |||||</marc:controlfield>
  <marc:controlfield tag="008">021113s1785 stknnn ||
    kneng</marc:controlfield>
  <marc:datafield tag="035" ind1=" " ind2=" ">
    <marc:subfield code="a">(DLC)13000786</marc:subfield>
  </marc:datafield>
  <marc:datafield tag="100" ind1="1" ind2=" ">
    <marc:subfield code="a">Kay, John,</marc:subfield>
    <marc:subfield code="d">1742-1826,</marc:subfield>
    <marc:subfield code="e">artist.</marc:subfield>
  </marc:datafield>
  <marc:datafield tag="520" ind1="0" ind2=" ">
    <marc:subfield code="a">Scottish cartoon shows a group of men conversing
      as balloons sail overhead. Possibly Scottish balloonist James Tytler and
      Italian balloonist Vincent Lunardi.</marc:subfield>
  </marc:datafield>  ...
```

(excerpt from LoC American Memory OAI Data Provider)

Search Protocols

- ❑ Z39.50 is the traditional remote search protocol for library systems.
 - ANSI/NISO/ISO standard
 - Comparatively complicated syntax/operation
 - Based on older standards (1998)
- ❑ ZING (Z39.50 International Next Generation) is the latest updated version.
 - SRW – Search/Retrieve for the Web
 - SRU – Search/Retrieve URL mechanism?
 - ❑ <http://myserver.com/myurl/searchRetrieve?query=dc.title=cat&maximumRecords=10&recordSchema=http%3a//www.loc.gov/mods/&sortKeys=title,dc&startRecord=1> (excerpt from ZING website)

SRW Request

```
SOAPAction: "searchRetrieve"

<SOAP:Envelope>
  <SOAP:Body>
    <SRW:searchRetrieveRequest
      xmlns:SRW="http://www.loc.gov/zing/srw/v1.0/">
      <SRW:query>(dc.author exact "jones" prox//5 title >=
"smith")</SRW:query>
      <SRW:sortKeys>/record/title,"http://www.loc.gov/zing/dcsche
ma/v1.0/",1,0,highValue
/record/datafield[@tag="100"]/subfield[@code="a"],"http://www.l
oc.gov/marcxml/" , , , "Smith"</SRW:sortKeys>
      <SRW:startRecord>1</SRW:startRecord>
      <SRW:maximumRecords>10</SRW:maximumRecords>

      <SRW:recordSchema>http://www.loc.gov/mods/</SRW:recordsSchema>
      </SRW:searchRetrieveRequest>
    </SOAP:Body>
  </SOAP:Envelope>
```

(excerpt from ZING website)

SRW Response

```
<SOAP:Envelope>
<SOAP:Body>
<SRW:searchRetrieveResponse xmlns:SRW="http://www.loc.gov/zing/srw/v1.0/"
  xmlns:DIAG="http://www.loc.gov/zing/srw/v1.0/diagnostic/">
<SRW:numberOfRecords>2</SRW:numberOfRecords>
<SRW:resultSetId>8c527d60-c3b4-4cec-alde-1ff80a5932df</SRW:resultSetId>
<SRW:resultSetIdleTime>600</SRW:resultSetIdleTime>
<SRW:records>
  <SRW:record>
    <SRW:recordSchema>http://www.loc.gov/mods/</SRW:recordSchema>
    <SRW:recordData> <?xml version="1.0" encoding="UTF-
8" ?> <mods xmlns:xlink="http://www.w3.org/TR/xlink"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://www.loc.gov/mods/"
xsi:schemaLocation="http://www.loc.gov/mods/
http://www.loc.gov/standards/mods/mods.xsd" >
  <titleInfo>
    <title>Sound and fury : the making of the punditocracy
  </title>
    <titleInfo>
    <name
type="personal" >
    <namePart>Alterman,
Eric.
  </namePart>
    <role>creator
  </role>
  </name>
  ...
  </SRW:recordData>
  <SRW:recordPosition>1</SRW:recordPosition>
</SRW:record>
  ...
```

(excerpt from ZING website)

Library Catalogue Systems

- ❑ Indexes MARC records.
- ❑ Provides OPAC (Online Public Access Catalogue) services to users.
- ❑ Provides library-library connections using protocols such as Z39.50.

- ❑ Examples: ALEPH (Ex Libris), Virtua (VTLS), Unicorn (Sirsi) ...

This document was created with Win2PDF available at <http://www.daneprairie.com>.
The unregistered version of Win2PDF is for evaluation or non-commercial use only.