LIBRARY CATALOGUES AND INFORMATION RETRIEVAL

PART 1: INTRODUCTION TO LIBRARY CATALOGUES & INFORMATION RETRIEVAL

In the world of libraries, there are two main types of information systems: those offering (usually automated) full-text indexing and those created manually, that is catalogues and indexes.

The full-text are just that: text only. Text-based retrieval is most effective on "medium-length" documents of related topics. The goal of storage and retrieval full-text is to locate only relevant documents, or high recall. High recall requires tuning system to the specific collection and skilled users. How might we "tune" the system and what do we know about the users, skilled or otherwise?

Attitudes towards how to tune the systems falls into two camps: hard core computer scientists who believe that all operations should be hidden from the end-user and increasingly refined mathematical manipulations create automatically-useful retrieval sets and others who believe integrating the user through the interface's controls or user profiles is a benefit because such models reintroduces the pragmatic aspect of language into IR … and generates data about how people actually think when using an IR system. The notion of how end-users cognitively process the IR interaction leads to many research endeavors. One is *information seeking behavior*; others include *cognitive science*, and the user's *mental models* of the system and of their own expectations.

Manually created catalogues and indices (catalogs and indexes) can be applied to all formats of material. To ensure the utility of these catalogues, there requires some kind of quality control of the metadata. This is, in fact, a goal of library school cataloguing classes. Likewise, it is believed that high recall requires tuning the system to the specific collection and through the use of skilled searchers (the librarians or trained end-users).

Descriptive Metadata:

Increasingly people – researchers and end-users – find that retrieval is improved when the metadata are somehow included in the IR system's behavior. Of course, this means building on the traditional idea of fielded searching (e.g., `author="Smith"`). As you know from reading about non-textual materials, the metadata standards in that arena are both rather well-developed but not entirely integrated into the searchable record. The Harvard University Visual Resources example suggests this. Aware of this, we can see that

the metadata are suitable for non-textual material searches, such as `type = "picture" and subject = "calligraphy"`.

Descriptive metadata can also be used with controlled vocabularies. This is evident from web-based retrieval. Consider searching a portal that requires authors to provide keywords (controlled vocabulary terms) combined with other CVs but from different arenas, such as HTML documents' metadata, e.g., `language="en"`.

Library catalogues vary depending on the organization's ability to maintain standards. Compare some library catalogues' interfaces and consider what differences there are and why.

http://www.simmons.edu/library
http://lib.harvard.edu/
http://www.loc.gov/rr/print/catalog.html

Origins of Library Catalogues

Many historians of librarianship start the discussion of the origins of catalogues from the writings of Sir Anthony Panizzi, Keeper of Books at the British Museum (1856-67). His *Ninety-One Rules* (1841) form the core of how librarians visualize catalogues: the bibliographic objective are (1) to bring together like items, and (2) to differentiate among similar ones.

Boston Athenæum's librarian, Charles Ammi Cutter, wrote in his *Rules for a Dictionary Catalog* (1874) that the goals of catalogue are:

1. To enable a person to find a book of which either the author, title, or subject is known
2. To show what the library has by a given author, on a given subject, or in a given kind of literature
3. To assist in the choice of a book as to its edition (bibliographically) or to its character (literary or topical).

Good ol' Melvyl Dewey when acting Librarian of Amherst College (1874) and later as founder of Columbia University's School of Library Service articulated goals of classification, his decimal system (Dewey Decimal) to cover the general fields of knowledge. Three digits numbers from 000 to 999 represent broad categories of knowledge and decimals to indicate specific topics within the broad categories; in other words a hierarchy. These rules made it possible to shelve books according to their primary subject content and to use sur-

rogates to create meaningful subsets quickly and achieve Dewey's goal of "library economy."

Technology

In the 19th century, the only items to be collected and organized were monographs and to this day there is a monograph-orientation in the processing of materials for storage and retrieval. For example, catalogues in the forms of books (Panizzi), index cards (Cutter) and the earliest online databases (Kilgour) were intended for the retrieval of monographs; later these same principles were applied to serials, maps, musical scores, and other non-text items

The creation of standards, catalogues, and then sharing materials in the catalogues (e.g., the *National Union Catalogue*, aka *NUC*) enable librarians to share records, combine records of the past with records created today, and allow the end-user and librarians to move between libraries. The cost involved in making catalogues can be high. One (wag?) writer suggests every catalogue record costs $100 to create. [But, you get what you pay for!] OCLC supposedly has invested several *billion* dollars in its >52,000,000 records.

Centralizing processing in OCLC means other libraries can copy some other library's more complete record, called *copy cataloguing*. Naturally, by adding only a library's holding code, a single record can be used by any number of other libraries. Without getting into too much detail, MARC records have a section, called the directory, that contains the rest of the record's metadata. These data are not really intended for human use, but are one of the earliest examples of data about data being included in a single record so that any (appropriately programmed) computer system could maintain the integrity of the record and extract its contents correctly for processing. There is, then, an *encoding* layer of a library catalogue. In addition, there are rules that define the fields and subfields, whether or not a field can be repeated, is optional, and so on. You'll recognize these as the *syntax* that is (or at least should be) created and stored in the system's *data dictionary*. Finally, there are rules that define the values of the fields and subfields, along with instructions for catalogues of what data to include and how to decide when choices are made. These are the *semantics* of the catalogue. These semantics appear both in the data dictionary and in the rules for humans, typically *Anglo-American Cataloguing Rules*, 2nd ed, revised (AACR2-Rev.).

AACR2-R is a joint publication of the American Library Association and the British Library Association. In it are all the rules for each category of material, guiding the librarian in what fields should be used and what data to include. Originally, the results of

what the librarian entered (if into a computerized system, such as OCLC) was used to generate library cards.

With the rules of machine readable cataloguing (MARC), computer-oriented rules were created for exchanging the format of catalogue records. The MARC Format includes encoding rules and syntax specification. The "MARC Catalogue" is the union of these two: a catalogue in MARC format where the content of each field follows AACR2.

AAR2 provides rules for (1) the choice of fields, (2) the content of the data that goes into each field, (3) the syntax of the data that go into each field. For an example, see AACR2 §22.1 (names). [Example to be added]

MARC Format was developed in the late 1960s as a tagging scheme for exchanging catalogue records on magnetic tape and remains the standard way today. MARC, however, does respond to technological changes: Unicode and XML are being integrated into MARC. Z39.50 remains useful.

Example:

Thoreau's text, *Walden; or, Life in the Woods*, is referred to casually as just *Walden*. The "official" citation style depends on one's standards, such as American Psychology Association (APA), a common citation style used in LIS: Thoreau, H. D. (1864). *Walden; or, Life in the woods*. Boston: Ticknor and Fields.

Here is the MARC record (from Harvard's holdings; notice non-standard fields that start with "H" and the HUL's unique system identifier):

```
FMT       BK
LDR       00596nam  2200205Ia 4500
001       007835722-5
005       20020606104010.5
008       850207s1864    mau           000 0 eng d
035 0     |a ocm11669001
040       |a ZCZ |c ZCZ |d HHG
090       |a PS3048 |b .A1 1854
100 1     |a Thoreau, Henry David, |d 1817-1862.
245 10    |a Walden / |c by Henry D. Thoreau.
260       |a Boston : |b Ticknor and Fields, |c 1864.
300       |a 357 p. ; |c 19 cm.
500       |a Title vignette (Thoreau's hut at Walden Pond)
752       |a United States |b Massachusetts |d Boston.
H01 0     |a BLZ2546
H03       |a MHBLZ25460HU
H05       |a PS 3048
SYS       007835722
```

The data in the column on the left are the tags. The values are on the right. These tags have "human" names, too: 245 = "Title Statement" aka Title proper; 260 = Pub-

lisher (Statement of Responsibility); 300 = Collation. In this example, there are no subject tracings (the 650 field).

Besides the data in the record, consider the author's name. On the title page, Thoreau uses his middle initial, but the catalogue record uses the official version of his name, Henry David Thoreau, stored in the Library of Congress (LC) name authority file (NAF). *Name authority files* are essentially catalogue records of humans – variants of their names, dates, epithets, the official version of the name, and resources used to create that name are brought together under a single access point, the name as sanctioned in the NAF. By creating an official version, the goals for a catalogue can be reached: to bring together all works by an author (regardless of variants). Here's an example of a name authority record. Notice the similarity to the monograph's MARC record.

```
LC Control Number:      n  87870182
        HEADING :       Arms, Caroline R. (Caroline Ruth)
    000             00907cz   2200205n  450
    001             4383796
    005             19890706143144.8
    008             70909n|acannaab |a aaa c
    010             __ |a n  87870182
    035             __ |a (DLC)n  87870182
    040             __ |a InU |c DLC |d DLC
    100             10 |a Arms, Caroline R. |q (Caroline Ruth)
    400             10 |w nna |a Arms, Caroline Ruth
    400             10 |a Arms, C. R. |q (Caroline Ruth)
    670             __ |a Arms, W.Y. Report on the performance
                       problems of the RLIN computer system, 1982:
                       |b t.p. (Caroline R. Arms)
    670             __ |a LC data base, 8/24/87
                       |b (hdg.: Arms, Caroline Ruth;
                       usage: Caroline R. Arms, C. R. Arms)
    670             __ |a Campus networking strategies, 1988:
                       |b CIP t.p. (Caroline Arms)
    670             __ |a Phone call to pub., 2/10/88
                       |b (Caroline Ruth Arms; studied at Oxford)
    670             __ |a Campus strategies for libraries and electronic
                       information, c1990: |b CIP t.p. (Caroline Arms)
                       data sheet
                       (b. 10-24-45)
    953   __ |a bz46 |b bd24
```

Subject Tracings

University and research libraries use the *Library of Congress Subject Headings* (LCSH) to create subject tracings that reflect topic, genre, and historicity. Here is the MARC record for Michael McCurdy's book *about* Thoreau. Notice the subject tracings.

```
FMT        BK
LDR           pam  2200325 a 4500
```

```
001       009310156-2
005       20040309145823.0
008       030709s2004    maua         000 0aeng
010       |a   2003054495
020       |a 1590300882 (acid-free paper)
035 0     |a ocm52729035
040       |a DLC |c DLC
043       |a n-us-ma
050 00    |a PS3048 |b .A1 2004
082 00    |a 818/.303 |a B |2 22
100 1     |a Thoreau, Henry David, |d 1817-1862.
240 10    |a Walden
245 10    |a Walden / |c Henry David Thoreau ; wood engravings by Michael McCurdy ;
          foreword by Terry Tempest Williams.
250       |a 1st Shambhala ed., [150th anniversary ed.]
260       |a Boston : |b Shambhala, |c 2004.
300       |a xiv, 303 p. : |b ill. ; |c 24 cm.
600 10    |a Thoreau, Henry David, |d 1817-1862 |x Homes and haunts |z Massachusetts
          |z Walden Woods.
650 0     |a Wilderness areas |z Massachusetts |z Walden Woods.
650 0     |a Natural history |z Massachusetts |z Walden Woods.
651 0     |a Walden Woods (Mass.) |x Social life and customs.
650 0     |a Authors, American |y 19th century |v Biography.
651 0     |a Walden Woods (Mass.) |v Biography.
650 0     |a Solitude.
SYS       009310156
```

This record demonstrates the use of subject headings and also hierarchical classification; the LCCN (call number) in the 050 field and DDC (Dewey Decimal Classification) in the 082 field. Creating and maintaining LCSH and LCCN is a never-ending task.

Towards Public Access:

Recall that MARC records were originally to generate library cards. It was only through the cards that the public gained access to the collection. The *view* of the data on-line was equally restricted; even reference librarians probably wouldn't have seen the entire MARC record, although cataloguers definitely would.

Some students are much too young to remember the early days of library catalogues (as I am, too!). Originally, libraries were leaders in making data available on a campus' central computing infrastructure. At the time, only dumb terminals permitted access to the mainframe, usually limiting searching to title and author, with limited Boolean searches.

As technology became more affordable and expectations changed, libraries began to convert their old records to MARC (*retrospective conversion*) and increase access to the catalogue by permitting off-site (out of the library) access.

In addition to home-grown or shared records (e.g., OCLC, RLIN), librarians

added secondary resources (usually commercial sources, such as SilverPlatter, Inspec, Med-Line, Chemical Abstracts, Dissertations International) and reference works (dictionaries, encyclopediæ), online, creating a multiplicity of interfaces, search syntaxes, and retrieval set elements.

As the technical backend improved (more fields became searchable, more integration between systems), interfaces were developed to permit more flexible searches, both fielded and full-text. Through some fits-and-starts (e.g., using Citrix), web interfaces were added, removing much of the proprietary nature of retrieval and minimizing the "cognitive load" of searching. Finally, some journals, especially scientific ones, move from publishing hard-copies and offered only online (electronic) subscriptions. In the same way, some scientific and business data (GIS, genome, census, marketing) were made available in their raw form.

Therefore, libraries moved away from dumb-terminals and locally-stored records towards *integrated library systems*, aka *library management systems*. Such systems integrate most of the library's technical functions (acquisitions, cataloguing, circulations, reference). End-users could search for materials, check their loan status and renew materials online. Some of the motivation for creating library management systems was from small enterprises. [One haughty computer scientist, like most of 'em, claims the reason is because libraries and these small vendors lack the capital and technical expertise to develop "modern digital libraries."]

Part 2 - IFLA MODEL AND THE CONCEPT OF A WORK; DUBLIN CORE AND XML

In Part 1, we were introduced to the founding principles of organization. What is being organized has been abstracted to the idea of an "information object" or a "work." Through efforts at international cooperation as well as the research efforts of people such as Smiraglia and Leazer, the view became popular that any physical object that could be information is a "work." A *work* is an underlying abstraction: a web site, an operating system, census data, a monograph's contents, *Hamlet* …, anything one could create. It is roughly equivalent to the concept of "literary work" in copyright law. [Some authors refer to this as the IFLA model.]

The abstraction is *realized through an expression*. We can have the work of Hamlet, but the written expression is Shakespeare's play, *Hamlet*. A film of actors acting the play becomes a work, too. The music played during the performance also becomes a work. A

virtual reality rendering of the characters reading the play are also a work. While most works have a single expression (e.g., a single book), the idea is to be able to collocate all manifestations of the same work. A *manifestation* is the expression given form in one or more ways. For example, all the editions of *Hamlet*.

An *item* is a copy of a manifestation, an actual single object. For example the notes created by the word processing program are an item; the printout you make from the computer file is also an item.

Enter the Web

It is interesting and relatively easy to create a digital library, IR system, etc., and use a test collection. The greatest difficulty when converting this test system to a real implementation is overcoming the challenges of *scalability*. A system that works effectively and efficiently with 10 users and 10,000 records may be a fiasco when used by 10,000 users and 10,000,000 records. Weibel (1995) noted this: "[automated indexes are most useful in small collections within a given domain. As the scope of their coverage expands, indexes succumb to problems of larger retrieval sets and problems of cross disciplinary semantic drift. Richer records, created by content experts, are necessary to improve search and retrieval." Aware of this, OCLC and others considered how to take the millions of already-created library records and use both the [potential] strength of the Web and IR techniques and the benefits of indexing to web-based materials. One result of the attempt to catalogue online materials is the *Dublin Core* (DC) [http://www.dublincore.org]

Like the MARC record above, DC consisted originally of a limited set of elements (15), intended for all types of genres of materials. All elements are optional and all are repeatable, something that is considered unwise in traditional database approaches. However, as is suggested by the following table, the elements have expanded to fulfill many needs. See `http://www.dublincore.org/documents/dcmi-terms/`

Metatags

It is informative to consider the tags themselves in a DC-tagged record. Here are some tags from the DC page listed above. Note the use of DC. This `DC.` is the *name-space*.

```
<meta name="DC.title" content="Dublin Core Metadata Initiative (DCMI)
     Home Page" />
<meta name="DC.description" content="The Dublin Core Metadata Initia-
```

```
    tive is an open forum engaged in the development of interoperable
    online metadata standards that support a broad range of purposes
    and business models..." />
<meta name="DC.date" content="2004-10-05" />
<meta name="DC.format" content="text/html" />
<meta name="DC.contributor" content="Dublin Core Metadata Initiative"
    />
<meta name="DC.language" content="en" />
```

You'll no doubt recognize the idea of elements and attributes and name-value pairs. In HTML for example, the <font> element has attributes of specific font names, e.g., face. The attribute name (face) accepts a value (e.g., "Verdana"): <font face="Verdana">. Moreover, these specific elements support *qualifiers*, e.g., "size=2"> to create <font face="Verdana" size="2">.

A DC qualifier for date, for instance, supports `created, issued, available`, and `valid` dates. [Can you think of when these would be valuable?] Note in the 2nd example DDC (Dewey Decimal Classification) and LCSH attributes.

```
Example: element qualifier
Example: Date
DC.Date.Created   1997-11-01
DC.Date.Issued    1997-11-15
DC.Date.Available 1997-12-01/1998-06-01
DC.Date.Valid           1998-01-01/1998-06-01
```

Value qualifier:

```
Example: Subject
DC.Subject.DDC        509.123
DC.Subject.LCSH       Digital libraries-United States
```

Similar to the idea of user's changing terms to broaden and narrow a search, the idea is that people who use metadata should be able to manipulate the data for more flexible searches. Wrote Lagoze (2001): "The theory behind this principle is that consumers of metadata should be able to strip off qualifiers and return to the base form of a property. ... This principle makes it possible for client applications to ignore qualifiers in the context of more coarse-grained, cross-domain searches." This concept is sometimes referred to by computer scientists as the "dumbing down principle."

This example shows the difference between a fully-qualified entry and a "dumbed-down" version. In the "dumbed-down" version, the date is a valid date and the DC.subject is a valid subject description:

```
Qualified version
DC.Date.Created   1997-11-01
```

```
DC.Subject.LCSH   Digital libraries-United States

Dumbed-down version
DC.Date           1997-11-01
DC.Subject        Digital libraries-United States
```

Complete Dublin Core Record:

OCLC: 45641346 *Entered:* DLC 2000-02-24
Not locked *System:* OCL 2001-01-03
Notify CNF of problem URLs in this record? no
No holdings in CNF *and no other holdings*

Title	Gore/Lieberman 2000
Title.alternative	Welcome to the Gore-Lieberman 2000 official campaign Web site
Title.alternative	Gore 2000
Title.alternative	Viva Gore Lieberman 2000
Identifier.LCCN	00530047
Identifier.URI	http://www.algore2000.com/
Type.OCLCg	Computer file
Type.AACR2-gmd	[computer file].
Contributor.nameCorporate	Gore/Lieberman, Inc.
Coverage.spatial.MARC21-gac	n-us---
Date.issued.MARC21-Date	2000-9999
Description.note	Title from home page as viewed on Nov. 1, 2000.
Description.summary	Presents information on U.S. Vice President Albert Arnold Gore, Jr. (b. 1948) and his presidential campaign, provided by Gore 2000, Inc. Contains biographical details about Gore, his wife, and family. Offers access to speeches, campaign news, and articles. Notes how to participate and contribute to the campaign.
Language.ISO639-2	eng
Language.ISO639-2	engspa
Language	In English and Spanish.
Publisher	Gore/Lieberman,
Publisher.place	Nashville, Tenn. :
Relation.requires	Mode of access: World Wide Web.
Subject.class.LCC	E840.8.G65
Subject.class.DDC	324.973
Subject.namePersonal.LCSH	Gore, Albert, • 1948-
Subject.topical.LCSH	Vice-Presidents • United States • Biography.
Subject.topical.LCSH	Presidential candidates • United States • Biography.
Subject.topical.LCSH	Presidents • United States • Election • 2000.
Subject.topical.LCSH	Political campaigns • United States.

From a computer scientists' perspective, emphasizing machine efficiency, DC records are considered "flat." All data about an item is held in a single DC record, including data about related items. This is convenient for access and preservation, but some information may be repeated, which can be a maintenance problem (the data can change one place but not be changed in this record – an absolute horror of data integrity). Linked records, on the other hand, store related data in separate records with a link from one to another. This is less convenient for access and preservation (can be; depends on the implementation) but at least

the data are stored only once and in one (hopefully recorded) place. The linked list idea is closer to vital issue of normal forms in relational databases.

Dublin Core and XML

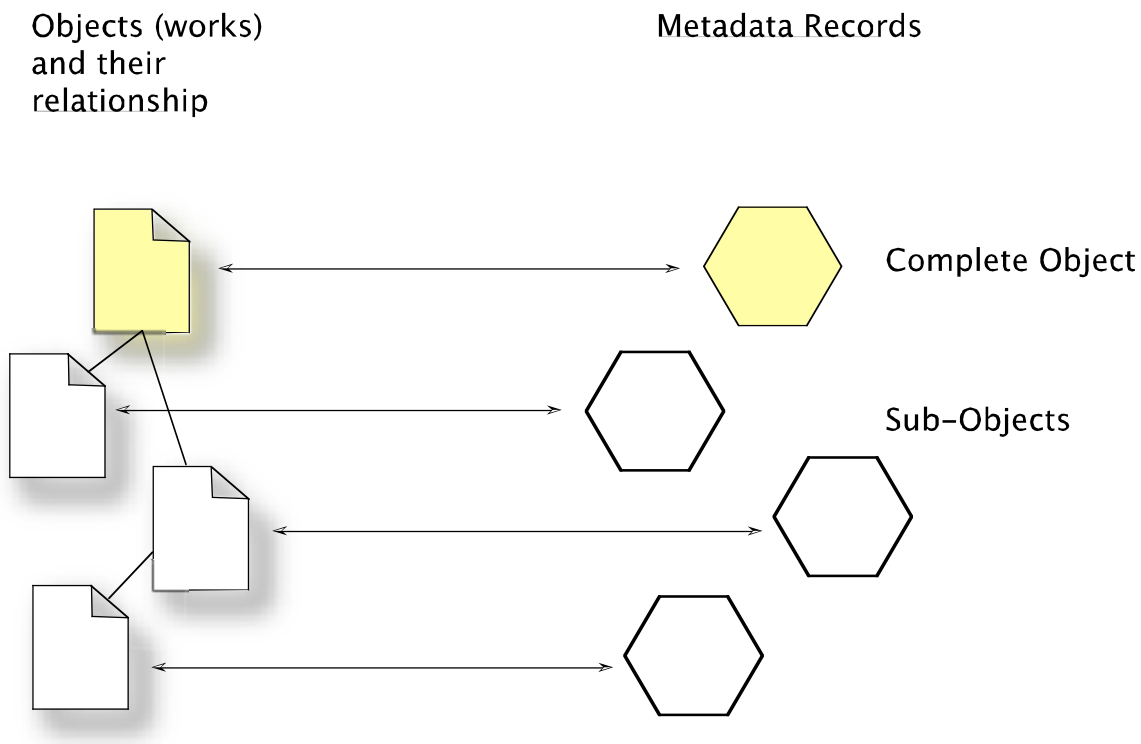
Here is a link to an original, "born-digital" article [not available in printed form, only online], David Levy's "Digital Libraries and the Problem of Purpose." Take a look: http://www.dlib.org/dlib/january00/01levy.html.

Having seen the original, note this surrogate of that electronic original in the form of a DC record with XML tags.

```
<title>Digital Libraries and the Problem of Purpose</title>
<creator>David M. Levy</creator>
<publisher>Corporation for National Research Initiatives</publisher>
<date date-type = "publication">January 2000</date>
<type resource-type = "work">article</type>
<identifier uri-type = "DOI">10.1045/january2000-levy</identifier>
<identifier uri-type =
"URL">http://www.dlib.org/dlib/january00/01levy.html</identifier>
<language>English</language>
 <rights>Copyright (c) David M. Levy</rights>
<relation rel-type = "InSerial">
            <serial-name>D-Lib Magazine</serial-name>
            <issn>1082-9873</issn>
            <volume>6</volume>
            <issue>1</issue>
      </relation>
```

Limits and Liabilities of Dublin Core and MARC:

Keep in mind the principles of retrieving all records/works and that works have potentially many and complicated relationships to other works. The "work" might consist of an article in a journal, a web page, an image, a newspaper article, a translated version of the article, and so on. If each of these items were a monograph in a [traditional] library system, then the [traditional] techniques of title proper, name authority records, etc., would be useful. Meanwhile, a metadata record that consists of a *single* object can miss the "sub-objects" represented by other works. This illustration may help make this point clear:



In the above illustration of objects and metadata records, what would happen if one of the objects (the dog-eared documents on the left) were updated, say from Version 1 to Version 2? Should Version 2's data be added to the original record, or should Version 2 have its own record? The new material traditionally has been noted in monographs as a new edition or somehow indicating that new material is available. This means there are two distinct works, which people often mistake as a new item.

If library catalogues have been successful as an information retrieval system (and they have been), why have they been successful? The basic operation is the same as our quadruple: matching a way a user describes an information need (query) against items in the collection. The success comes from the use of precise language to describe items (standards, such as LCSH and AACR2) combined with trained and experienced users, formulating queries (and these include reference librarians). If library records are being converted to DC and XML, why not use DC to index and search the web?

One reason is that the lack of technologies that were evident when the Internet first became popular have been greatly fulfilled. [Note that these methods provide good precision at the expense of low recall.] Moreover, seekers do not know anything about cata-

logue construction standards and have limited training. Finally, there's the issue of cost: it is impossible to index every important web site and the rate of change (documents being updated or changed, dropped or added) could require frequent reindexing. [This is the *current* concern. Who knows about tomorrow!]

Further reading: Baker, T. (1998, Dec.). Languages for Dublin Core. *D-Lib Magazine*. Available: http://www.dlib.org/dlib/december98/12baker.html

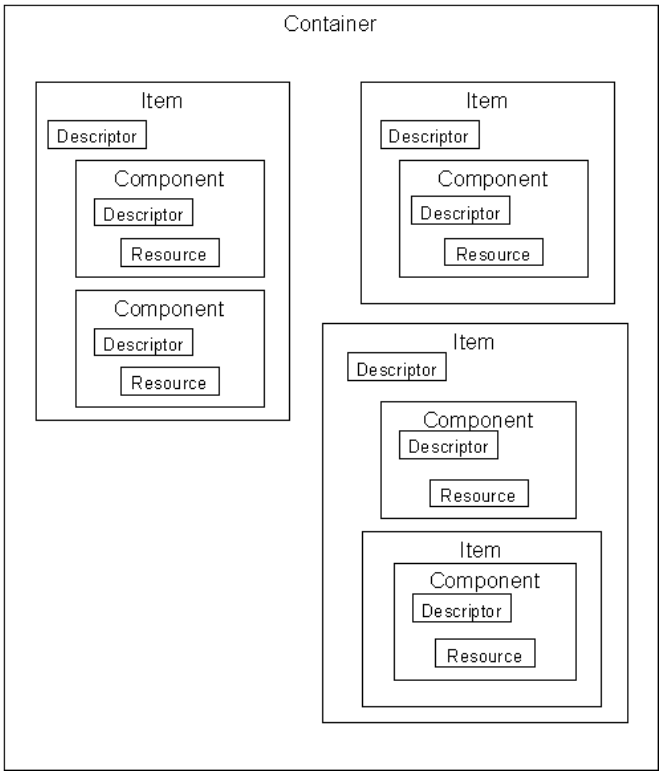
See also foreign language and mixed language retrieval, called CLIR, or *cross-language information retrieval*.

Part 3 - SCHEMA AND PACKAGING OF LIBRARY DATA

The data in a library system, both text and visual, are packaged. The very packaging is, of course, a form of information useful in the storage and retrieval of data in IR.

One standard is the *Metadata Object Description Schema* or MODS http://www.loc.gov.standards/mods/. A standard for sound and images is MPEG 21 http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm.

Here is an MPEG container. Notice how the concepts of a work appear (as "container") and specific items appear (as "items"). Inside each component notice the descriptor and resource. Programmers will have noticed the similarity to class definitions.



The purpose of that graphic is to emphasize to you the idea of *reusability*: how can old data be used and reused in ways that can be standardized, shared, computationally efficient and effective for end-users?

Reusing MARC:

Consider the MARC record. When MARC was created, the encoding schema for computers creating these records was ASCII. ASCII is just fine, of course, but encoding standards have modernized. Instead of countless national industrial standards, such as KOI-8, JIS (Japanese Industrial Standard), IIS (Indian Industrial Standard), Win-char-1225 (Windows Character Set 1225), and so on, some computer programming languages, like Java, and some operating systems, the Macintosh, and begrudgingly Windows XP, have adopted Unicode, a 16-bit character encoding schema, as the internal representation of data, that is, how the data are stored on disk. MARC records that use Unicode as their encoding standard become useful throughout the world, regardless of script or language.

Similarly, instead of original MARC, the notion of reusability has led to converting the MARC format to an open standard, XML. The result is MARCXML, aka "Marc 21 XML" (http://www.loc.gov/standards/marcxml/] for direct conversion of xml tagging and to the MODs, a subset of MARC with data clean-up.

MARCXML

This schema retains the semantics of MARC; it is intended to be simple and flexible. The fields are treated as elements with the tag as an attribute; and indicators treated as attributes. Subfields are treated as sub-elements with the subfield code as an attribute. [Compare this to traditional RDMS columns and rows.]

Data conversion often means a loss of data. This project emphasizes the *lossless conversion* of MARC to XML – and back! (referred sometimes oddly and one trusts jokingly as "roundtripability"). MARCXML also inherits the benefits of pure XML: data can be presented differently depending on need through an XML *stylesheet*, XSL, the data can be *validated* (as all XML records should be), and the record is *extensible* (extendable).

MODS.

From the official website for MODS, Metadata Object Description Schema, we read: "The Library of Congress' Network Development and MARC Standards Office,

with interested experts, has developed a schema for a bibliographic element set that may be used for a variety of purposes, and particularly for library applications. As an XML schema, the 'Metadata Object Description Schema' (MODS) is intended to be able to carry selected data from existing MARC 21 records as well as to enable the creation of original resource description records. It includes a subset of MARC fields and uses language-based tags rather than numeric ones, in some cases regrouping elements from the MARC 21 bibliographic format. MODS is expressed using the XML schema language of the World Wide Web Consortium. The standard is maintained by the Network Development and MARC Standards Office of the Library of Congress with input from users." (http://www.loc.gov/standards/mods/).

```
<mods>
<titleInfo>
    <title>Sound and fury :</title>
    <subTitle>the making of the punditocracy /</subTitle>
</titleInfo>
<name type="personal">
    <namePart>Alterman, Eric</namePart>
    <role>
          <roleTerm type="text">creator</roleTerm>
    </role>
</name>
<typeOfResource>text</typeOfResource>
<originInfo>
    <place>
         <placeTerm type="text">Ithaca, N.Y</placeTerm>
    </place>
    <publisher>Cornell University Press</publisher>
    <dateIssued>c1999</dateIssued>
</originInfo>
<language>
        <languageTerm authority="iso639-2b"
                      type="code">eng
        </languageTerm>
</language>
</mods>
```

Automatic Extraction of Catalog Data

The creation of manual records by trained catalogues results in high quality records. The complaint from managers (and oddly from computer scientists) is that cataloguing is expensive and time-consuming. In response, there are movements to automate entirely catalog data from DC-tagged records for web pages. The idea is that such records would be produced quickly and at almost no cost. The liability is the quality. [No problem –

have a human go through the record!, it's been proposed.]

Researchers at the University of Bath UKOLN have created a software package (http://www.ukoln.ac.uk/metada/dcdot/) called **DC-dot**. This application creates a skeleton DC record from clues in the web page and there's a GUI for cataloguers to edit the resulting record. Notice that there are difficulties when processing a collection of records vs. a single page.

Using DC-dot on this class's (single) homepage generates the following:

```
<link rel="schema.DC" href="http:/purl.org/dc">
<meta name="DC.Title" content="LIS531 – Information Retrieval">
<meta name="DC.Subject" content="Welcome to LIS531, Information
Retrieval {cut text} Posted July 21, 2005.">
<meta name="DC.Publisher" content="Simmons College">
<meta name="DC.Data" scheme="W3CDTF" content="2005-07-21">
<meta name="DC.Type scheme="DCMIType" content="Text">
<meta name="DC.Format" content="text/html">
<meta name="DC>Format" content="5781 bytes">
<meta name="DC.Identifier" con-
tent="http://web.simmons.edu/~benoit/LIS530/">
```

Note that the web page's html tag `<title>` becomes DC.Title. The DC.Publisher is the owner of the IP address where the page is found. DC.Subject is a list of headings and noun phrases presented for editing. DC.Date is taken from the Last-Modified field in the http header. DC.Type and DC.Format are taken from the MIME type of the http response. DC.Identifier was supplied by the user as input.

Collection-level metadata: [see Jenkins and Inman]

Several of the most difficult fields to extract automatically are the same across all pages in a web site. Therefore, consider making a collection record manually and then combining it with the automatic extraction of other fields at the item level. For example, the our class, LIS531h, collection-level metadata are:

```
<meta name="DC.Publisher" content="Simmons College">
<meta name="DC.Creator" content="Gerald Benoit">
<meta name="DC.Rights" content="Gerald Benoit, 2005">
```

Now let's return to Levy's article and see what DC-dot extracts automatically.

D.C. Field	Qualifier	Content
title		Digital Libraries and the Problem of Purpose
subject		*not included in this slide*
publisher		Corporation for National Research Initiatives
Date	W3CDTF	2000-05-11
Type	DCMIType	Text
format		text/html

format		27718 bytes
identifier		http://www.dlib.org/dlib/january00/01levy.html

Collection-level Data:

publisher		Corporation for National Research Initiatives
type		article
type	resource	work
relation	rel-type	InSerial
relation	serial-name	D-Lib Magazine
relation	issn	1082-9873
language		English
rights		Permission is hereby given for the material in D-Lib Magazine to be used …

Combined item-level record [DC-dot plus collection-level]

D.C. Field	Qualifier	Content
title		Digital Libraries and the Problem of Purpose
publisher		(*) Corporation for National Research Initiatives
date	W3CDTF	2000-05-11
type		(*) article
type	resource	(*) work
type	DCMIType	Text
format		text/html
format		27718 bytes
relation	rel-type	(*) InSerial
relation	serial-name	(*) D-Lib Magazine
relation	issn	(*) 1082-9873
language		(*) English
rights		(*) Permission is hereby given for the material in D-Lib Magazine to be used …
identifier		http://www.dlib.org/dlib/january00/01levy.html

(*) indicates collection-level metadata

Manually Created Record Version:

D.C. Field	Qualifier	Content
title		Digital Libraries and the Problem of Purpose
creator		(+) David M. Levy
publisher		Corporation for National Research Initiatives
date	publication	January 2000
type		article
type	resource	work
relation	rel-type	InSerial
relation	serial-name	D-Lib Magazine
relation	issn	1082-9873
relation	volume	(+) 6
relation	issue	(+) 1
identifier	DOI	(+) 10.1045/january2000-levy
identifier	URL	http://www.dlib.org/dlib/january00/01levy.html
language		English
rights		(+) Copyright (c) David M. Levy

(+) entry that is not in the automatically generated records

Collection-Level Metadata

Finally, we should compare the metadata extracted automatically by DC-dot and consider each level accordingly (collection-level, combined item-level, and manual records). For web pages, IR works better by automatic indexing, rather than automatic extraction of metadata followed by indexing of metadata.

It is interesting to note that D-Lib magazine was edited by a librarian. The pages are created with the user in mind, the quality of tagging (and writing) are maintained to be useful to humans with little regard for making the site rank highly by search engines. It is noteworthy because this site lends itself to successful automatic indexing.

Other web pages, especially marketing ones, are *not* successfully indexed automatically and there is little consistency in the page layout and tagging. In a study of the Bush and Gore presidential web sites (Belew, 2000) it's been suggested that these pages were intended as marketing more than anything else and constructed by public relations specialists with the goal of ranking highly by search engines. Such sites are difficult to index automatically. The significance of these phenomena are not fully understood.

However, automated extraction of metadata from video sequences is actually effective. [See the work by Informedia.]

A collaborative team of members of Cornell University's Human Computer Interaction group and Prof. Liz Liddy at Syracuse University has created "Metatest." This project compares effectiveness *as perceived by the end-user* of indexing based on comparing (1) manually-created Dublin Core records, (2) automatically created DC (but higher quality than DC-dot), and full-text indexing. The preliminary results suggest little difference between them as perceived by end-users. The reasons for this, too, are obscure, but one wouldn't be unwise to suspect the research model may be faulty.