

Manual cataloging and indexing

Min-Yen KAN

*heavily drawn from Lancaster (98)
*Indexing and Abstracting in Theory
and Practice*



Objectives of the Library

Ranganathan (1957)

- Books are for use
- Every reader his book
- Every book its reader
- Save time of the reader
- The library is a growing organism



Mesopotamian Catalogs

- Mesopotamians kept track of their tablets with a list of their incipits:

- What is it?
A poem?

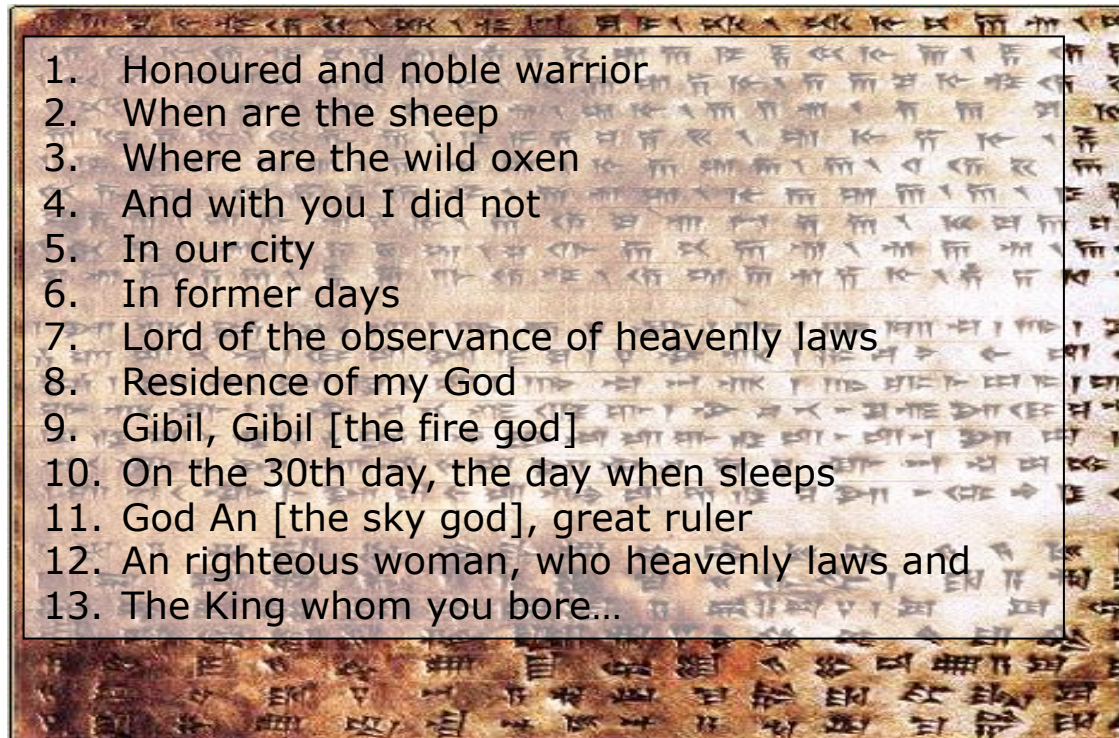


Image from: i-cias.com/e.o/cuneiform.htm



Some Definitions

1. (Subject) Indexing

- Assigning index terms to represent a document
- Assists in document retrieval

2. Classification

- Assigning a label to a document to assist in organizing that information
- Not necessarily semantic labels

Linguistic Relativity

- Also known as Sapir-Whorf hypothesis
- A loose definition:
Our language to some extent determines the way in which we view and think about the world around us.
- An example: time
 - Tomorrow = day after today
 - ð (“bukra”) = some point in the future
- The result?
 - No single best representation
 - Every representative offers a bias
 - Many AI researchers reject / ignore this notion



Steps in Subject Indexing

1. Conceptual analysis
 - Determine “aboutness”
 - Computational approaches: $TF \times IDF$
2. Translation
 - Expressing the concepts as index terms



Conceptual analysis

- Generic: What is it about? What's the main content
 - *e.g.*, The History of Sociology
- Specific: Why has it been added to our collection? What aspects will our users be interested in?
 - *c.f.*, "Every reader his book"

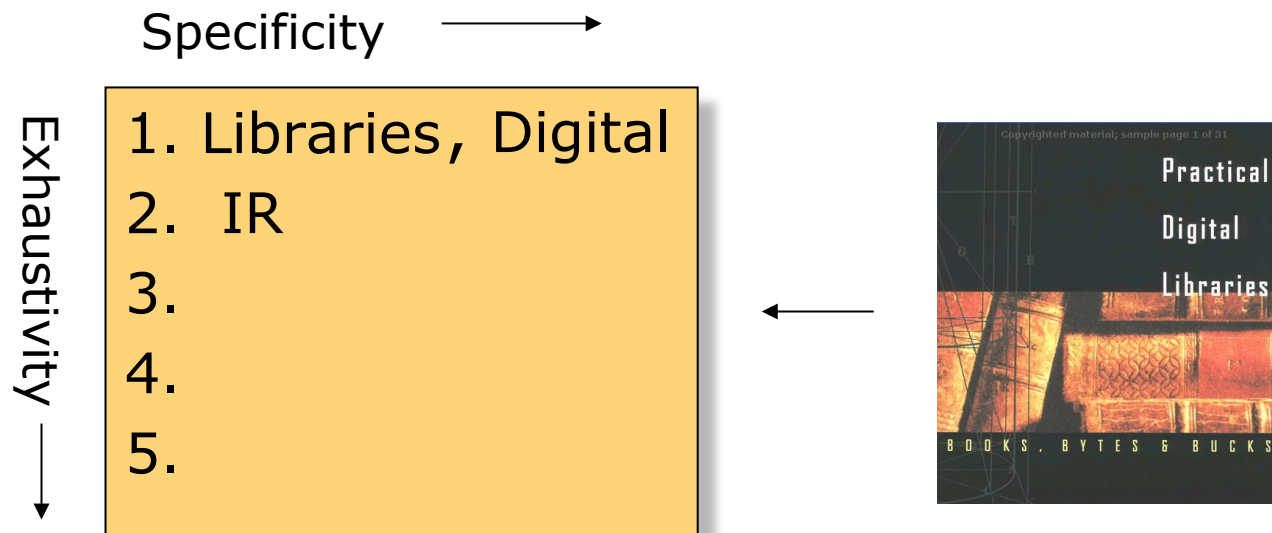
Thus, organizations index differently

- Different subjects (specialty, general interest)
- Different materials (own materials, 3rd party)



Index terms

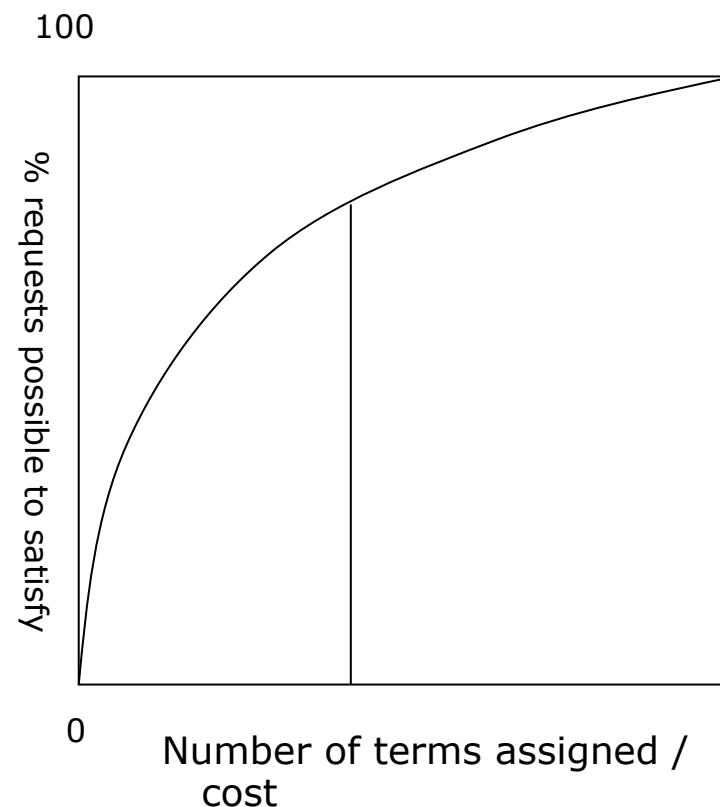
- Assigning content terms increases one or the other axis



- How do we choose index terms then?

Number of index terms in record

- Long (Exhaustive)
 - Gives good recall at cost of precision
 - Few records fit in the UI
 - Hard to figure out which are main aspects
- Short (Selective)
 - Gives good precision at cost of recall
 - Less work
- In practice: offer levels of indexing for tasks
 - Index Terms
 - Abstract



- Modified from Lancaster (98), pg 27.



Translation

- Extraction: use terms directly from the source itself
- Assignment: use terms from an outside source.
 - Usually from a controlled vocabulary.



Controlled vocabulary objectives

1. Control / suggest synonyms, pick an authoritative term
 - Especially for entities: people (maiden names to married names), places (St. Petersburg)
2. Distinguish among homographs (e.g., mercury, turkey)
3. Link terms with their relationship (is-a and all others (associative))

Difficulties in Naming Authorities

- People

- Use most common name:
Dr Seuss
Not Theodore Seuss Geisel



- Geographic Names

- Use latest name:
Namibia
Not Zaire

-- Examples from AACR 2

- Data must be constantly updated to provide users with best access points – not an easy job

Controlled vocabulary usability

- Good structure to find the appropriate term
 - Standard fields in an CV:
 - USE/UF: Use instead / Use For (authoritative)
 - BT/NT: Broader / Narrower Term in terms of hierarchy
 - RT: Related Term (Associative Term)
- Applied by experienced personnel
 - A large vocabulary can be hard to map to

Question: What to do if the controlled vocabulary has no term for the concept to be indexed?

Controlled vocabulary examples

General CVs

- Sears List of Subject Headings
 - More general divisions, not intended for research libraries
 - Geared towards general subdivisions
- Library of Congress Subject Headings (LCSH)
 - Comprehensive, very large, over five volumes



Domain-specific CV

- Medical Subject Headings (MeSH)
 - Byproduct of indexing the NLM
- Art & Architecture Thesaurus (AAT)
 - Object, images, architecture, styles
- ERIC Thesaurus
 - Educational materials (journals, lesson plans and computer files)



Classification



Objectives of classification

- Uniqueness
 - Be able to fetch a specific resource given a call number
- Notational Permanence
 - (Seldom) have to reorganize/reassign labels
 - (e.g., paradigm shift in mathematics)
- Comprehensiveness
 - Can successfully classify most things

Inventory

- Serendipity
 - Collocate related subjects together
- Ease of Use
 - Ways of resolving ambiguities
 - (e.g., given religious architecture and Egyptian architecture, where does an article on the architecture of Egyptian temples go?)

Access



Types of classification

- **Enumerative**
 - Produce an alphabetical list of subject headings, assign numbers to each heading in alphabetical order
- **Hierarchical**
 - Recursively divides subjects hierarchically, from most general to most specific
- **Faceted** (analytico-synthetic):
 - Analytic: Divides subjects into mutually exclusive orthogonal facets
 - Synthetic: Combine facets to get a new class

- From Taylor (92)



Dewey Decimal Classification

- Divide knowledge into ten classes
 - Recursively divide these categories into ten (or fewer classes)
 - Assign another digit
 - What type of classification scheme is it?
- 000 Generalities
 - 100 Philosophy & psychology
 - 200 Religion
 - 300 Social sciences
 - 400 Language
 - 500 Natural sciences & mathematics
 - 600 Technology (Applied sciences)
 - 700 The arts
 - 800 Literature & rhetoric
 - 900 Geography & history

Top level of DDC

ACM Classification scheme

- Four-level tree
 - 3 coded levels and
 - a fourth uncoded level)

- 16 General Terms

- | | |
|--------------------|---------------------|
| 1. Algorithms | 9. Management |
| 2. Design | 10. Measurement |
| 3. Documentation | 11. Performance |
| 4. Economics | 12. Reliability |
| 5. Experimentation | 13. Security |
| 6. Human Factors | 14. Standardization |
| 7. Languages | 15. Theory |
| 8. Legal Aspects | 16. Verification |

- | |
|---|
| H. Information Systems
H.0 GENERAL
H.1 MODELS AND PRINCIPLES
H.2 DATABASE MANAGEMENT (E.5)
H.3 INFORMATION STORAGE AND RETRIEVAL
H.4 INFORMATION SYSTEMS APPLICATIONS
H.5 INFORMATION INTERFACES AND PRESENTATION (e.g., HCI) (I.7)
H.m MISCELLANEOUS
I. Computing Methodologies
I.0 GENERAL
I.1 SYMBOLIC AND ALGEBRAIC MANIPULATION
I.2 ARTIFICIAL INTELLIGENCE
I.3 COMPUTER GRAPHICS
I.4 IMAGE PROCESSING AND COMPUTER VISION
I.5 PATTERN RECOGNITION
I.6 SIMULATION AND MODELING (G.3)
I.7 DOCUMENT AND TEXT PROCESSING (H.4 , H.5)
I.m MISCELLANEOUS |
|---|

Colon Classification

- Raganathan proposed 5 basic facets (**PMEST**):
 - **Personality** – the subject matter
 - **Material**
 - **Energy** – process or action
 - **Space**
 - **Time**
- Each facet would have its own classification schedule
- String together notation to get classification number

Example:

The design of wooden furniture in 18th century America



The **design** of **wooden furniture** in 18th century **America**



The **D21** of **W12 F290** in R2 U20



D21:W12:F290:R2:U20



Classification Maintenance

- DDC and LCSH ~~are~~ were centralized systems
- Nowadays, rely on a distributed approach to update
 - Either hierarchically determined authorities
 - Or arbitration of conflicts
 - Think CVS and source control systems



To think about...



**Calvin says:
What, think now?!?**

- Now that we have free-text searching, do you feel controlled vocabularies are still necessary or not? What do you feel their impact will be in the future of the digital library?
- How would you improve the ACM classification scheme? How to deal with legacy schemes?
- Booksellers also need to use classification to shelve books. Which type of classification do you think booksellers use? Would you make any adaptations to the classification schemes shown today?

The Dewey Decimal Classification

000	General Works
100	Philosophy & Psychology
200	Religion
300	Social Sciences
400	Language
500	Natural Sciences & Mathematics
600	Technology
700	The Arts
800	Literature
900	Geography & History

The Dewey Decimal Classification

000	General Works
100	Philosophy & Psychology
200	Religion
300	Social Sciences
400	Language
500	Natural Sciences & Mathematics
600	Technology
700	The Arts
800	Literature
900	Geography

Classification

000	General Works
100	Philosophy & Psychology
200	Religion
300	Social Sciences

Soda Break

See ya!





Metadata creation and management*

*Parts of this lecture come from Lilian Tang's lecture material at the Univ. of Surrey



What is metadata, anyways?

- Data about data
 - From the DB community
- “Cataloging or indexing information that [information professions] create to arrange, describe, and otherwise enhance access to an information object”
 - -- Gilliland-Swetland (1998)
- “Data that describes attributes of a resource, characterize its relationships, support its discover and effective use and exist in an electronic environment”
 - -- Vellucci (1998)



Outline

- What is Metadata?
 - Some Frameworks →
- Packaging Metadata
 - Warwick Framework
- Structural Metadata
- Hidden Web Metadata
 - OAI
 - SDARTS
- Crosswalking and Automated Extraction

Metadata formats

- HTML Metadata
- AACR2 / TEIH / MARC / Z39.50
- Dublin Core



Types of metadata

- Administrative
- Structural
- Descriptive
- Intellectual Property
- Use

Metadata that describes the accessibility and ownership of the object

- Copyright
- Encryption techniques
- Access information
- Publisher

Metadata of the disclosure

partially examined this in the last class.

• Origin

Metadata related to the level and type of use of the object

- Exhibition records
- Use and user tracking

Editorials records

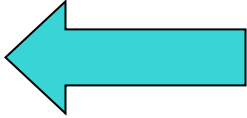


• Location

- Annotations by users
 - Finding aids
- (also metadata)



Metadata attributes

Adapted from Baca (1998)

Attribute	Characteristic
Source 	<ul style="list-style-type: none">• Internal – File name, parameters for digitization• External – Rights, cataloging records
Nature	<ul style="list-style-type: none">• Lay metadata – Personal filing systems• Expert metadata – MARC records
Status	<ul style="list-style-type: none">• Static – Title, creation date• Dynamic – User access / transaction logs
Semantics / Structure	<ul style="list-style-type: none">• Controlled – MARC• Uncontrolled – Free text
Creation 	<ul style="list-style-type: none">• Automatic – Keyword indices• Manual – Written by an individual• Semi-Automatic – Tool for controlled vocabulary
Level 	<ul style="list-style-type: none">• Collection• Item / subitem

These key factors influence all other aspects (source ↔ creation)

- Can you think of other relationships?



Data types: MIME

- Multipurpose Internet Mail Extensions
(text/plain, img/jpg application/msword)
 - Simple format, pre-web
 - Can code an unofficial type using x-subtype prefix (e.g., audio/x-pn-realaudio)
 - Application tag: need to use an application to handle this data
 - Wild success shows a simple system is best:
 - Good for adoption / authoring
 - Good for common denominator

Complex objects and granularity

- DOI identifier records: multiple versions of a single document (hi res / low res)
- Syntax should be mirrored in reference metadata

Structural

Image

Shot

Clip

Movie

Trilogy

Sentence

Paragraph

Text

Chapter

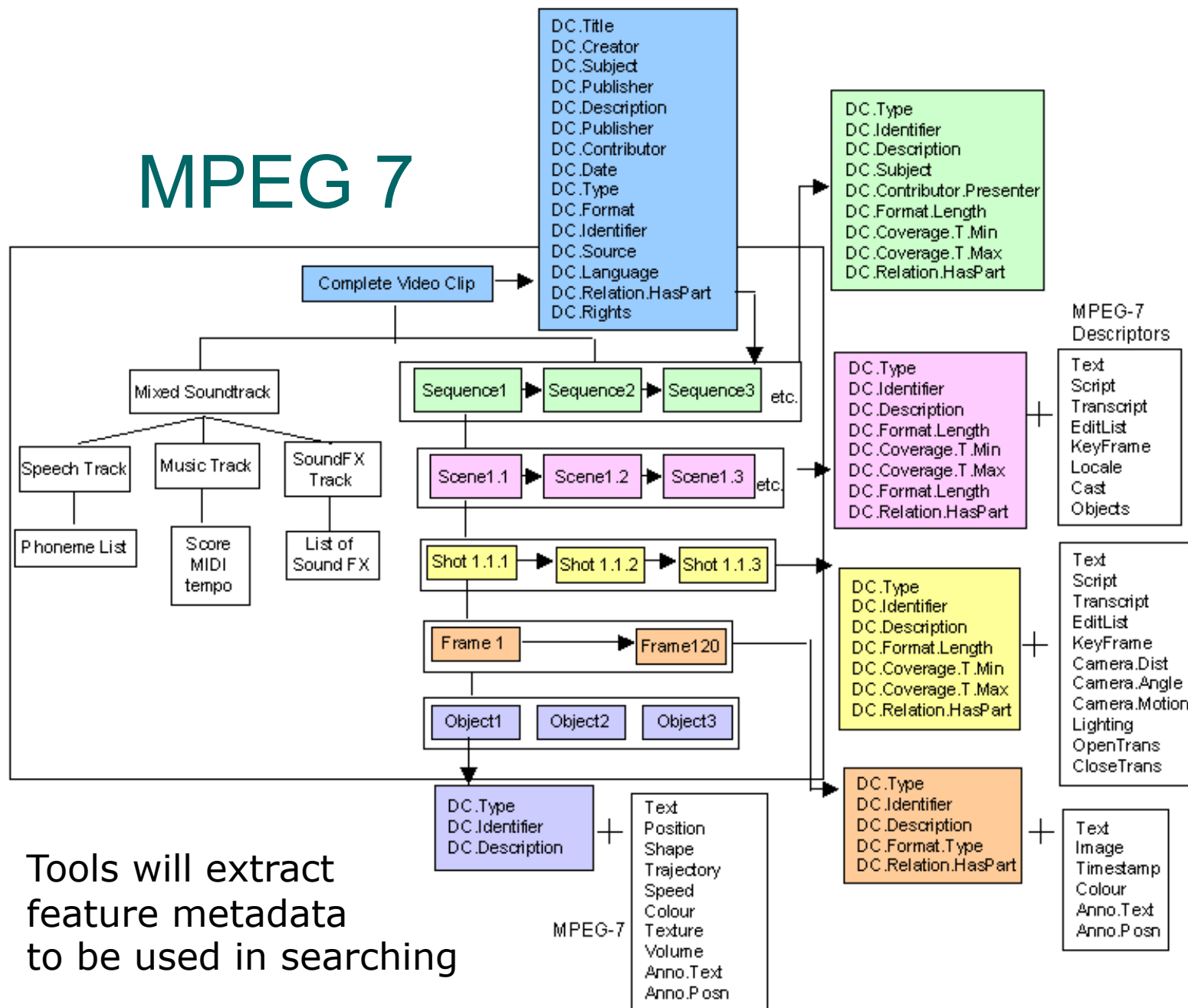
Book

Reference

Structural

Reference

MPEG 7



Tools will extract feature metadata to be used in searching



Audio/visual metadata

- Based again on how people search (The Potato Eaters)
 - I'm looking for a picture of a group.
 - I'd like it to be a family group.
 - This family should be doing something that would be typical for a family, like sitting around a table with food in front of them, look grateful for what they have to eat.

Facet analysis is a good approach

- Objective (“of”)
- Subjective (“about”)

Visual:

- Who / What is shown?
- What is happening?
- Why? How? When? Where?

Audio:

- What is recorded?
- Who has been recorded?
- What's happening?
- Why? When? Where?



HTML Meta tags

```
<HTML><HEAD>  
  <META NAME="attribute" VALUE="value">  
</HEAD>... </HTML>
```

- Not regulated or controlled
- You can add your own tags
- Only certain ones parsed by finding aids (e.g., GoogleBot)
- Many tags use other metadata formats

Examples include

- Content Type
- Keywords
- Description
- Language
- Author
- Version



MARC / AACR 2 / TEIH

- **Machine Readable Cataloging**
 - Standard for encoding cataloging data (bibliographic and authority)
 - Standoff Annotation (External)
- **Anglo American Cataloguing Rules 2**
 - Set of rules used for collecting bibliographic data and for formulating access points (for authors, titles, subjects, related works, etc.)
 - Regulates format and number of access points
- **Text Encoding Initiative Header**
 - Header, similar to <HEAD> in HTML
 - Is located within the document (Internal)
- **Z39.50**
 - Protocol for clients to ask queries of servers

Librarians use AACR2 / TEI to devise values for fields to be encoded by MARC (external) or in TEIH (internal). This data is accessible by users using the Z39.50 protocol.



Data Types

Used to describe the different types of (complex) objects in the digital library

- Structural facets of documents



Dublin Core Elements

A common denominator set of metadata attributes used for interoperability. Has recommended values for some fields.

```
<dc:title>Metadata</dc:title>
<dc:author>Kan, Min-Yen</dc:author>
<dc:description>Digital Library course module on metadata</dc:description>
<dc:type>InteractiveResource</dc:type>
<dc:subject>XML, RDF, metadata, Dublin Core Metadata</dc:subject>
<dc:format>appn/powerpoint</dc:format>
<dc:identifier>http://www.comp.nus.edu.sg/~kanmy/courses/6210_2003/m4-
metadata.ppt</dc:identifier>
<dc:rights>Copyright 2003, Min-Yen Kan</dc:rights>
```

- Besides Title, Creator, Publisher, Contributor, 11 other fields:
- 5. Subject
 - Subject, expressed as keywords, key phrases or classification codes that describe a topic of the resource.
 - value from a controlled vocabulary or formal classification scheme.



Dublin Core Elements (Con' t)

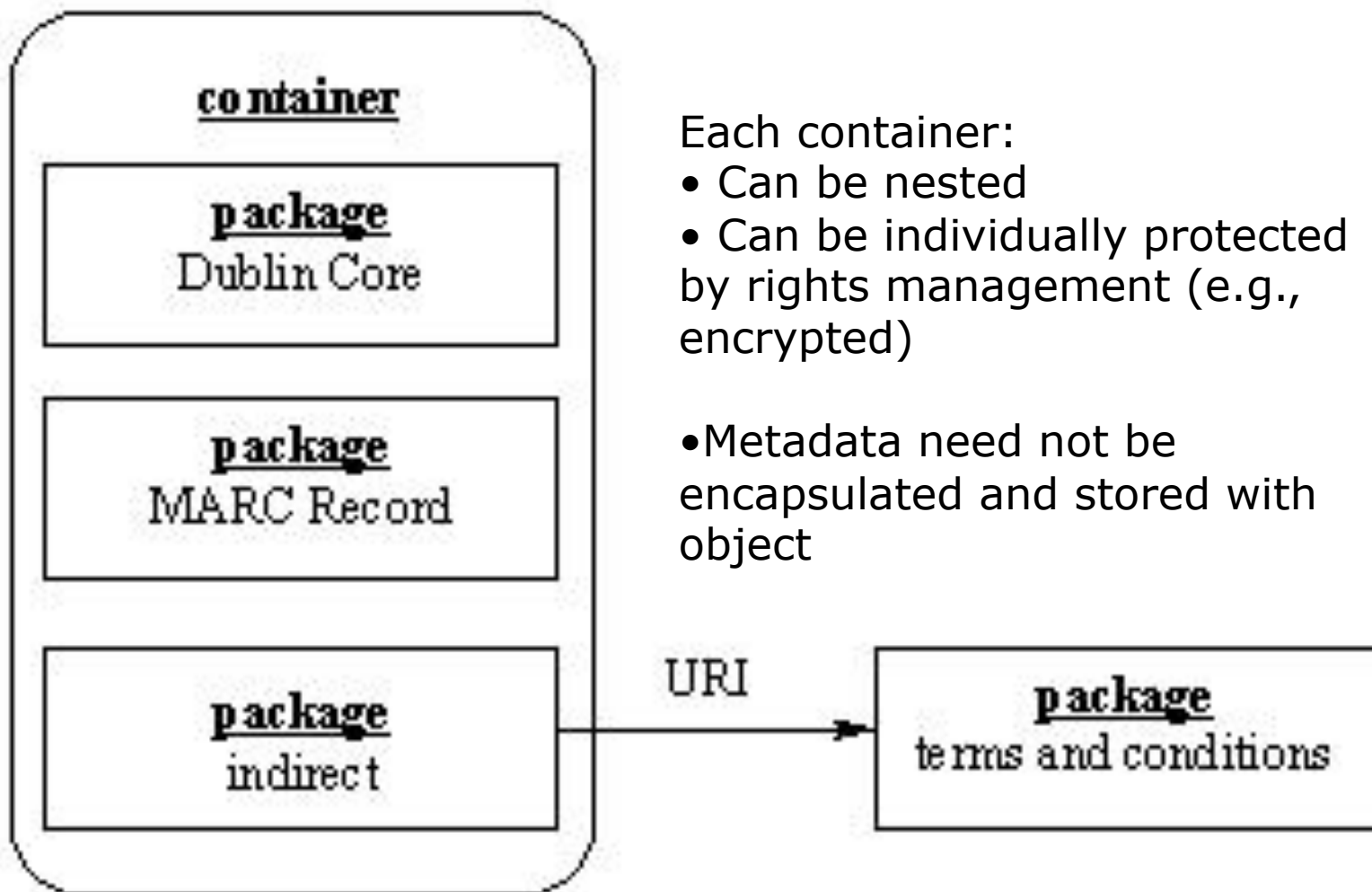
6. Description
 - An account of the content of the resource.
 - (e.g., an abstract, ToC, graphical representation of content or a free-text account of the content)
7. Date
 - creation or availability of the resource
 - ISO 8601 (e.g., YYYY-MM-DD)
8. Type
 - The nature or genre of the content of the resource
 - value from a controlled vocabulary (e.g., DCMI Type Vocabulary)
9. Format
 - Media-type or dimensions. Also, identifies the software, hardware, or other equipment needed to display or operate the resource.
 - value from a controlled vocabulary (e.g., MIME)
10. Identifier
 - Unambiguous reference to the resource within a given context.
 - Use a formal identification system, (e.g., URI, DOI, ISBN)



DC Elements (Con' t)

11. Source
 - Reference to a resource from which the present resource is derived (e.g., past edition)
12. Language
 - Language of the intellectual content of the resource.
 - use RFC 3066 with ISO639 (e.g., “en-GB”)
13. Relation
 - Reference to a related resource
14. Coverage
 - The extent or scope of the resource (e.g., location, time period)
 - value from a controlled vocabulary
15. Rights
 - Statement of copyright or a reference to one
 - If absent, no assumptions may be made

Warwick Example



STARTS: A Metasearching Protocol

- Defines:
 - Query language
 - Results format
 - **Metadata for the collection**
- No specified transport layer or implementation
- Built to assist metasearchers.

Why does the metadata help metasearchers?

- Hint: Ranking documents

Query Operators

Example of metadata:

Stemming = no

of docs = 20,000

...

Diabetes → TF:12, DF: 4

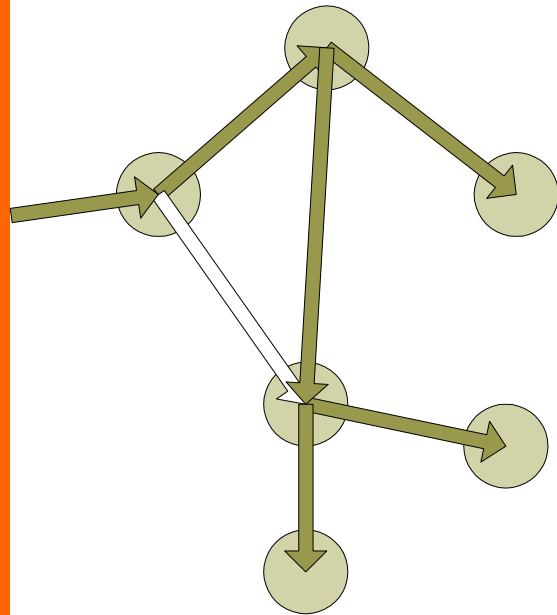
XML → TF:1200, DF:750

...

Frequency of Collection

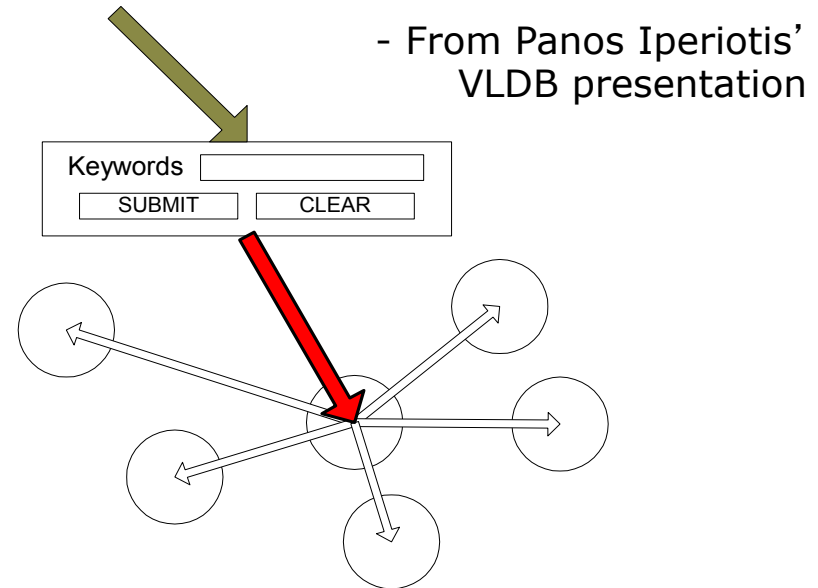
Distributed Search? Why?

“Surface” Web vs. “Hidden” Web



● “Surface” Web

- Link structure
- Crawlable
- Documents indexed by search engines



● “Hidden” Web

- No link structure
- Documents “hidden” in databases
- Documents not indexed by search engines
- **Need to query each collection individually**

Hidden Web: Examples

- From Panos Iperiotis' VLDB presentation

PubMed search: [diabetes]

→ **178,975** matches

PubMed is at <http://www.ncbi.nlm.nih.gov/PubMed>

Google search: [diabetes site:www.ncbi.nlm.nih.gov]

→ **119** matches

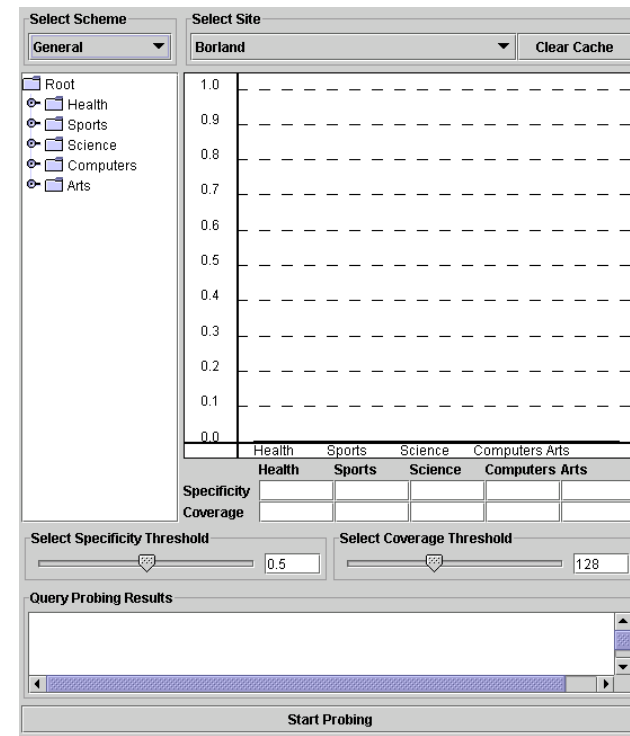
Database	Query	Matches	Google
PubMed	diabetes	178,975	119
U.S. Patents	wireless network	16,741	0
Library of Congress	visa regulations	>10,000	0
...

Query Probing

- Idea: Send different queries to categorize data
- Demo Time!
(if it works)

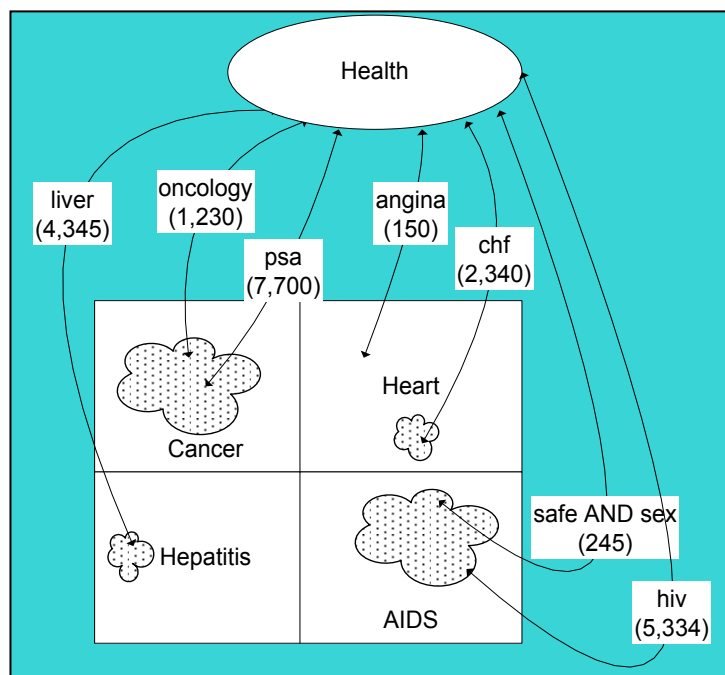


Pikachu says
“pika pika!”



	Health	Sports	Science	Computers	Arts
Specificity					
Coverage					

Focused Probing: Sampling



Sampling proceeds in rounds:
In each round, the rules associated with each node are turned into queries for the database

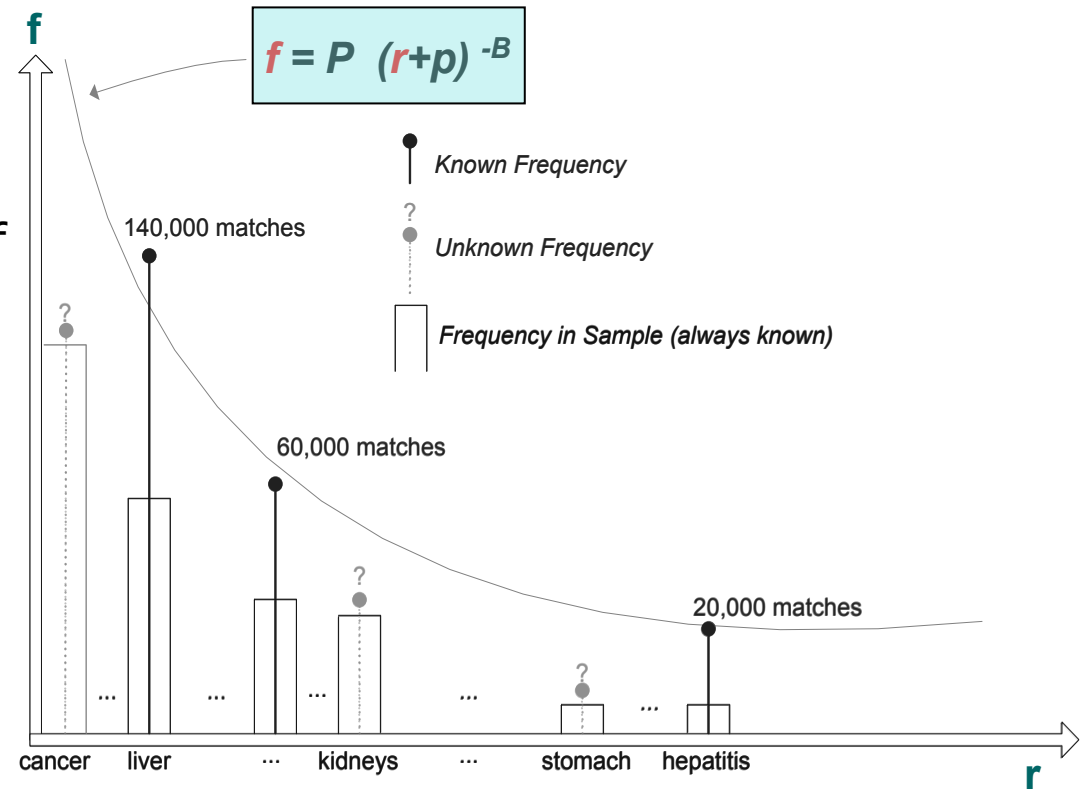
- Transform each rule into a query
- For each query:
 - Send to database
 - Record number of matches
 - Retrieve top- k matching documents
- At the end of round:
 - Analyze matches for each category
 - Choose category to focus on

Output:

- Representative document sample*
- Actual frequencies for some “important” words*

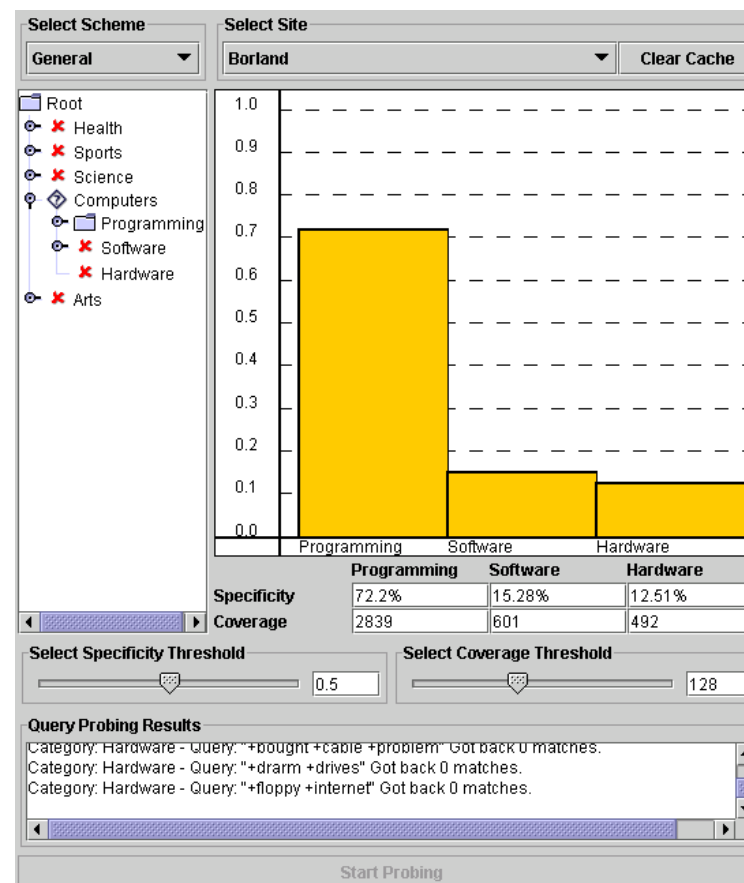
Adjusting Document Frequencies

- We know ranking r of words according to document frequency in sample
- We know absolute document frequency f of some words from *one-word queries*
- Mandelbrot's formula connects empirically word frequency f and ranking r
- We use curve-fitting to estimate the absolute frequency of **all words** in sample



Focused Probing

- Algorithm:
 - Send general queries to determine high level category
 - Send progressively more specific queries to determine mid- and lower-categories



Automatic Extraction of Metadata

- Rule-based scripts (fragile):

- **DC Dot** Demo:



- <http://www.ukoln.ac.uk/cgi-bin/dcdot.pl>

- Still heavily cited and used!

- Wrapper induction: localized extraction

- Define a local context and features to match and extract

- Text classification: classification

- Use features over the entire document to determine classification.

Open Archives Initiative (OAI)

A low-barrier interoperable standard for the dissemination of content

- In principle, not tied to a specific purpose
- Note: open in terms of open architecture, not necessarily) free



○ Protocol for **M**etadata **H**arvesting

- Defines standard for advertising metadata in a repository.
- Standard packages for harvesting have been defined.

○ DP 9

- A standard for exposing metadata to web crawlers as web pages.

MeURLin

Classification of URLs to the Open Directory Project



http://www.onlineshawnee.com/stories/072901/ent_shelton.shtml

- Doesn't require webpage, just address
- About 1/2 - 1/3 as accurate as full words approaches
- Uses scalable segmentation and expansion techniques



Crosswalking

The transfer of metadata from one format to another

- Retrofitting = updating old metadata to a newer format
- Aids accessibility and discovery
- Complementary to OAI / SDARTS (which are centralized approaches)
- Mostly done manually by specialists
 - CS work to be done here!



Reference

- Dublin Core Tool List
 - <http://www.lub.lu.se/tk/metadata/dctoollist.html>
 - And many others
- The Getty Research Institute
 - <http://www.getty.edu/research/institute/>
- Crosswalking
 - <http://www.ukoln.ac.uk/metadata/interoperability/>

Summing Up



- Metadata authoring highly intricate but two complementary purposes
 - Inventory
 - Access (what we care more about)
 - Uses CV standards (licensing drawback)
- Automated approaches have promise ...
 - To access and annotate more data
 - But generally needs re-work, or NLP post-processing to make data fit standard
- Questions?!?