Modern Information Retrieval

Chapter 16

Library Systems

Environment in the Library
Online Public Access Catalogues (OPACs)
Document Databases
IR in Organizations

Environment in the Library

- Web provides ubiquitous informational repository
  - but chaotic and unstructured
  - many of the sources may be questionable regarding accuracy, reliability, completeness
  - results for query "information retrieval" on Google
    - wikipedia article on IR
    - 1979 book on IR by van RIjsbergen
    - homepages of IR journals
    - major AI association
    - first edition of "Modern Information Retrieval"
  - that is, user can only be satisfied within the limitations of the Web

Environment in the Library

- Many searches do not admit a Web solution
  - diagnosis or treatment of a disease
  - surgery procedure
  - case law and precedents
  - patent rights
- Libraries provide IR services for
  - collections, books, journals, digital materials
  - all systematically acquired and organized
- There will always be closed repositories in the corporate world with vast repositories of data

Environment in the Library

- **Hybrid Library**
  - digital and traditional hard copy materials co-exist
  - brought together in an integrated information service
  - accessible on-site, like a traditional library, and remotely through the Internet or within a local network
- Philosophical assumption
  - hybrid libraries are about organized access, rather than local collections

Environment in the Library

- Main challenge for libraries
  - to create integrated and seamless access to local and remote resources
  - users, who now also search the Web, expect access to be free, convenient, and simple
  - in a library, subscriptions and licenses govern use of material
  - seamless Web experience is hard to provide

Environment in the Library

- **Library retrieval systems**
  - provide access to a range of databases
    - Online Public Access Catalogues (OPACs)
    - commercial abstract and indexing services
    - electronic journals
    - collections of e-books
    - special (digitized) collections
    - institutional repositories
  - provide (or try to provide) a unified "look and feel" across the library, even if many distinct search systems compose the library
  - facilitate hyperlinking among the disparate search systems that compose the library

OPACs

- **Catalogue**
  - access point for materials held by library
- **Integrated Library System (ILS)**
  - system that manages all library catalogues and collections
- **OPAC**
  - early on, libraries used card catalogues
  - later, followed on microfilms and microfiche forms
  - in the 1970's, online catalogues were implemented
    - initially, they had very limited functionality
    - in the 1980's, true online public access catalogues (OPACs)
    - today, OPACs are the main component of an ILS

OPACs

- OPACs used standardized record formats (MARC)
  - minimal subject information (title, a few headings)
- Three generations
  - first generation: known-item finding tools through search by author, title, control number
  - second generation: increased search technology with access by subject headings, keywords, boolean queries
    - problems included failed searches, navigational confusion
    - enhancements represented large investments for a library
  - third generation: focus on open systems architectures, improved GUI, support for Z39.50 and Dublin Core, hypertext links, java programming, ranked results sets
    - problems include slow pace of development
    - failure to match trends in the Web

OPACs and Bibliographic Records

- Libraries use standardized systems for cataloguing and classifying materials
  - **Anglo-American Cataloguing Rules:** to describe materials
  - **Library of Congress or Dewey Decimal Classification:** to assign subject codes
  - **Library of Congress Subject Headings:** to assign subject descriptors
- For sharing information, they rely on centralized bibliographic utilities to
  - lower the cost per unit to catalogue materials
  - broaden access through shared databases
  - facilitate the sharing of materials

Centralized Bibliographic Utilities

- Broaden access through shared databases
- Facilitate the sharing of materials
- **Online Computer Library Center (OCLC)**
  - used by 69,000 libraries in 112 countries and territories
  - union catalogue of collections over 10,000 libraries
- **WorldCat**
  - 125 million bibliographic records
  - 1.3 billion library holdings
  - opened to the public in 2006 as *worldcat.org*

MARC

- MARC is the **Machine Readable Cataloguing Record**
  - underlines cooperation among libraries
  - provides support to distinct online catalogues
  - data format that implements national and international standards
    - **Information Interchange Format (ANSI Z39.2)**
    - **Format for Information Exchange (ISO 2709)**
  - variations in the world: **USMARC**, **UKMARC**

MARC Sample Record

MARC 21 Bibliographic Format

Full Level Record - Motion Picture

Network Development and MARC Standards Office
Library of Congress

This example can be identified as a record for projected material by code g in Leader/06, and more specifically as a motion picture by code m in field 007/00. This record illustrates the use of several MARC data elements to describe an archival motion picture, including: the use of character positions 09-/22 in field 007, and multiple occurrences of fields 007, 300, and 541 for the several versions of the motion picture being described. Other noteworthy data elements include: the use of field 017 (Copyright or Legal Deposit Number); field 040, subfield $e (Description conventions); field 257 (Country of Producing Entity for Archival Films); and field 510 (Citation/References Note).

LDR		*****cgm##22*****#a#4500
001		<control number>
003		<control number identifier>
005		19920513133548.3
007		mr#bf##dnnartnnac198607
007		mr#bf##dnnbdtnnac198607
007		mr#bf##dnnaetnnac198607
008		870505 s1918#### xxu055 ####### #####ml ####d
017	##	**$a**LP12321**$b**U.S. Copyright Office
040	##	**$a**<organization code>**$c**<organization code>**$e**amim
245	00	**$a**=M'liss /**$c**Pickford Film Corp. ; supervised and directed by Marshall A. Neilan ; photoplay by Frances Marion.
257	##	**$a**U.S.
260	##	**$a**United States :**$b**Artcraft Pictures Corporation,**$c**1918.
300	##	**$a**5 reels of 5 on 2 (1988 ft.) :**$b**si., b&w ;**$c**16 mm.**$3**ref. print
300	##	**$a**5 reels of 5 on 2 (1988 ft.) :**$b**si., b&w ;**$c**16 mm.**$3**dupe neg.
300	##	**$a**5 reels of 5 on 2 (1988 ft.) :**$b**si., b&w ;**$c**16 mm.**$3**arch pos.
500	##	**$a**Copyright: Famous Players-Lasky Corp.; 18Apr18; LP12321.
500	##	**$a**Originally released in 35 mm..
500	##	**$a**Based on a story by Bret Harte.
508	##	**$a**Photographed by Walter Stradling ; art director, Wilfred Buckland.
510	4#	**$a**New York times film reviews,**$c**5-6-18.
510	4#	**$a**Variety film reviews,**$c**5-10-18.
510	4#	**$a**Moving picture world,**$c**v. 36.l, p. 894, 897, 1043.
511	1#	**$a**Mary Pickford (M'liss), Theodore Roberts (Bummer Smith), Thomas Meighan (Charles Gray), Charles Ogle (Yuba Bill), Tully Marshall (Judge Joshua McSnaggley), Monty Blue (Mexican Joe), Val Paul (Jim Peterson), Winnifred Greenwood (Clara Peterson)

MARC Record

- Composed of three parts
  1. fixed length leader (24 characters)
  2. record directory showing the 3-digit tag for each field
  3. data containing the fields and subfields
     - subfields indicated by codes, such as $a
- Example: field 260
  - contains publication information
  - may contain subfields for place, publisher, and date
- MARC is more useful for *known-item search*
  - field 650 might be added for subject headings
  - field 505 might be added for table of contents
  - field 520 might be added for summaries or annotations

IR from the ILS

- In OPAC, only metadata search (no full text)
  - known-item search: find full information on a specific item
  - subject search: usually limited to title and subject headings
- Academic Library OPAC—University of British Columbia

Information Retrieval from the ILS

- Keyword search provides ranked output
- When subject is limited, ranking might be poor
- In title search, in-field match might not be provided
  - full match is done from left to right
  - a search for "Information Retrieval" will not match "Modern Information Retrieval"
- Endec ProFind Guided Navigation, from North Carolina State University, provided a fix
  - export MARC records to be indexed by search engines
  - provides them relevance ranking and faceted search
  - index updated daily

UC Bib Services Task Force

- Key recommendations
  - Provide users with direct access to an item
  - Provide recommender features
  - Support customization/personalization
  - Offer alternative actions for failed or suspect searches
  - Offer better navigation of large sets of search results
  - Deliver bibliographic services where the users are
  - Provide relevance ranking and leverage full-text
  - Provide better searching for non-Roman materials [**?**, pages 3–4]
- For re-architecting the OPAC
  - create single catalog UI for all university collections
  - support searching across all bibliographic space

Integrating the Hybrid Library

- OPAC allows searching materials in the library collection, but offers limited functionality to other resources
- Consider a search for an e-journal title
  - answer is link to external Web site of a publisher
  - because e-journals are bundled, further links available
  - if license specifies individuals, additional authentication
- Library Web site usually offers link to Google Scholar
  - federated search to a variety of sources
  - includes theses and dissertations
  - not comprehensive, since relies on material on the Web

OPACs and End Users

- End users use OPAC only infrequently
  - underling record structure (MARC) is detailed and complex
  - organizational structures (LCSH, LC) are not intuitive
- OPAC search
  - most common form is subject search
  - failures in topical search are well documented
  - study of transactional logs at Nanyang Technical University
    - average query length is 2.82 terms
    - only 12% of searches used Boolean operators
    - almost half the queries returned "zero results"

ILS: Vendors and Products

- Library System Vendors
  - **SirsiDynix**
    - largest ILS vendor
    - systems in 4,000 libraries
    - current systems are Unicorn and Symphony
  - **Innovative Interfaces**
    - second largest ILS vendor
    - academic, public, special and school libraries
  - **Libris**
    - large company
    - targeting academia libraries and consortia
    - ALEPH and Voyager systems

ILS: Vendors and Products

- Systems developed within research projects and implemented in academia
  - **MELVYL**, **Okapi**, **Cheshire**
- Open Source Software for Libraries
  - **Evergreen**
    - developed by Georgia Public Library Service for use in its PINES network
    - now used in hundreds of libraries across US and Canada
    - support and development company, Equinox Software, founded by developers of Evergreen
    - other implementations include **Pines** and **Sitka**

ILS: Vendors and Products

- Open Source Software for Libraries
  - **Koha**
    - developed in New Zealand
    - first open-source ILS
    - entreprise-class ILS

ILS: Vendors and Products

- Google Book Search offers deeper search of the full text
- Libraries now adding new Web 2.0 features
  - RSS feeds
  - user tagging and reviews
  - federated search
  - navigational aids
  - relevance ranked output
  - better visual appeal
  - **AquaBrowser** is a good example of this trend

Document Databases

- Libraries offer access to wide range of external materials
  - electronic databases
  - bibliographic databases: contain citations and abstracts
  - document or full-text databases: contain full text articles
    - search remote sites
    - license and mount them locally
- History of ILS
  - 1950: demonstration of the use of computer for library search
  - 1964: first system of library search by the National Library of Medicine (NLM), using batch processing
  - 1970's: Lockheed's DIALOG system implemented for NASA

Bibliographic and Full-text DBs

- Abstracting and indexing tools in printed form first available in the 19th century
- Professional organizations, commercial firms, government bodies served as publishers
- First databases concentrated on the sciences
  - Chemical Abstracts
  - Biological Abstracts
  - Engineering Index
- Humanities and social sciences products became soon available
  - Historical Abstracts
  - PsycINFO

Bibliographic and Full-text DBs

- Today, all printing and indexing products are online
- Many are available only in electronic form
- Some-well known databases on DIALOG
  - **Chemical Abstracts**
    - bib records for world-wide literature of chemistry
    - over 20 million records
    - weekly updates of 18,000 records
  - **MEDLINE**
    - bib records for materials in the life sciences with emphasis on biomedicine
    - about 15 million records from 4,300 journals
    - yearly updates of 400,000 records

Bibliographic and Full-text DBs

- Some-well known databases on DIALOG
  - **NY Times**
    - full text of New York Times from 1980 to present
    - over 2.8 million records
    - daily updates
  - **PsycINFO**
    - bib records for psychology, behavioral and social sciences
    - over 2.6 million record from 1,700 journals
    - weekly updates of 1,500 records

Content of Database Records

- Sample record: **BIOSIS Previews**

2/9/1 DIALOG(R)File 5:Biosis Previews(R) (c) 2006 The Thomson Corporation. All rts. reserv.
0015888673 **Biosis No.**: 200600234068
Maximum body size among insular Komodo dragon populations covaries with large prey density

Author: Jessop Tim S (Reprint); Madsen Thomas; Sumner Joanna; Rudiharto Heru; Phillips John A; Ciofi Claudio
Author Address: Zool Soc San Diego, Ctr Reprod Endangered Species, San Diego, CA 92112 USA **USA
Author E-mail Address: timj@uow.edu.au
Journal: Oikos 112 (2): p 422-429 FEB 2006 **2006**
ISSN: 0030-1299
Document Type: Article
Record Type: Abstract
Language: English

Abstract: This study documents variation in maximum body size of Komodo dragons (Varanus komodoensis) among the four extant island populations in Komodo National Park and compares an indirect measure of deer density, the major prey item for large dragons, to differences in maximum body size among islands. The largest 15% of dragons from the large islands of Komodo and Rinca were significantly longer and heavier than the largest 15% of dragons on the small islands of Gili Motang and Nusa Kode. There was a 33% difference in snout vent length (SVL) between dragons found on Komodo and those found on Gili Motang, with mass varying by more than four-fold. Density of deer pellet groups between islands ranged from 5.86 +/- 0.75 groups per transect on Gili Motang to 20.73 +/- 1.02 groups per transect on Komodo Island. Maximal dragon SVL and mass was highly positively correlated with this index of deer density. Low prey density on the two small islands could constrain body size via energetic constraints. At present we can not deduce if insular body size variation has arisen through genotypic or phenotypic mechanisms.

Descriptors:
Major Concepts: Terrestrial Ecology--Ecology, Environmental Sciences; Biogeography-- Population Studies
Biosystematic Names: Cervidae--Artiodactyla, Mammalia, Vertebrata, Chordata, Animalia; Sauria-- Reptilia, Vertebrata, Chordata, Animalia
Organisms: deer (Cervidae)--prey; Varanus komodoensis (**Komodo dragon**) (Sauria)
Common Taxonomic Terms: Artiodactyls; Mammals; Nonhuman Mammals; Animals; Chordates; Nonhuman Vertebrates; Reptiles; Vertebrates

Geographical Name: Komodo National Park (Indonesia, Asia) (Oriental region); Gili Motang (Indonesia, Asia) (Oriental region); Nusa Kode (Indonesia, Asia) (Oriental region); Rinca (Indonesia, Asia) (Oriental region)
Miscellaneous Terms: genotype; phenotype; extinction; prey density; body size variation; maximum body size; snout vent length

Content of Database Records

- Sample record: **Historical Abstracts**

9/9/175 DIALOG(R)File 39:Historical Abstracts (c) 2005 ABC-CLIO. All rts. reserv.
1683680 54-6856
SETTLING THE CANADIAN COLONIES: A COMPARISON OF TWO NINETEENTH-CENTURY LAND COMPANIES.
Browde, Anatole
Business History Review 2002 76(2): 299-335.
Document Type: ARTICLE
Abstract: Compares the performance of two British land companies-the **Canada** Company and the British American Land Company-chartered to sell land and encourage emigration to the colonies of Upper and Lower **Canada** during the 1820's-40's. The **Canada** Company was not only fiscally responsible but also fully aware of Canadian conditions. In addition, it engaged in strategic planning for its operations. The British American Land Company, on the other hand, was badly managed. It undertook little in the way of strategic planning, was managed solely for the benefit of the proprietors, and poorly understood Canadian conditions. The **Canada** Company accomplished its mission of facilitating emigration to **Canada** after 1815, while the British American Land Company failed in this endeavor. Based on record books for the Bank of England, **Canada** Company Papers in the Ontario Archives (Toronto), Colonial Office records in the Public Record Office (Kew), British American Land Company and **Canada** Company papers in the National Archives of **Canada** (Ottowa), and other primary and secondary sources; 89 notes. (H. M. Friedman)
Descriptors: Great Britain| **Canada|Emigration|Land (sale of)|Canada Company|British American Land Company|1820's-1840's**
Historical Period: 1820D 1830D 1840D 1800H
Historical Period (Starting): 1820's
Historical Period (Ending): 1840's

Historical Abstracts (Dialog® File 39): (c) 2005 ABC-CLIO. All rights reserved.

Database Producers & Vendors

Database Producers	Database Vendors
Design database structure	Create search software
Collect in-scope literature	License databases from producers
Enter bibliographic information in standard form	Standardize (as possible) record structure
Abstract (or edit authors' abstracts)	Mount databases, creating inverted indexes
Index with (usually) controlled vocabulary	Update databases as appropriate (daily, weekly, monthly)
Generate file updates at regular intervals	Provide documentation for searchers
Market backfile and updates to vendors	Market to clients
	Provide service and training to client base

Some Database Vendors

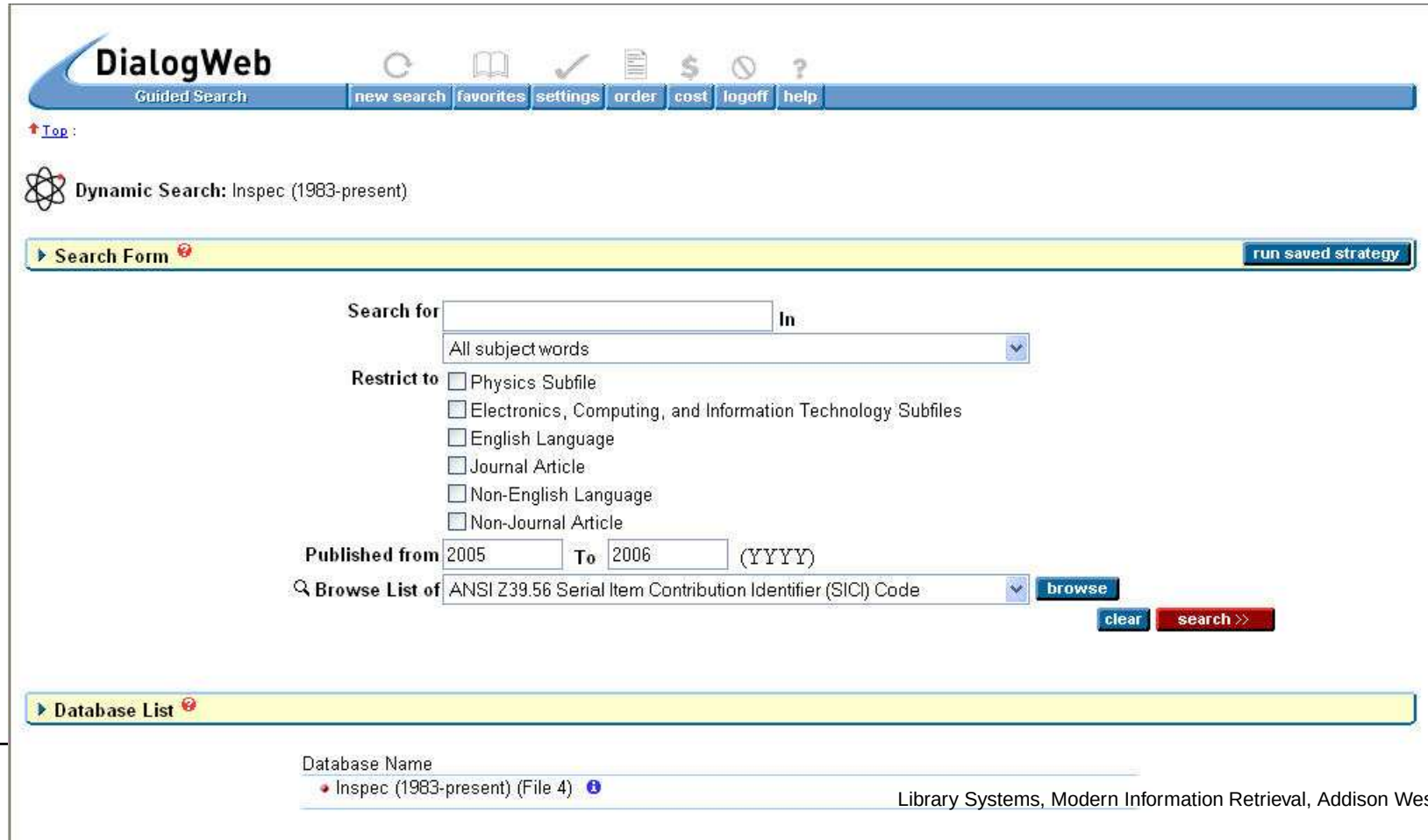
- **DIALOG**
  - leader in information for research in science, engineering, business and intellectual property
  - more than 600 databases
  - 1.5 billion unique records
- **LEXIS-NEXIS**
  - full-text databases to the legal and business community
  - over 5 billion searchable documents
  - more than 40,000 legal, news and business sources
  - access reliability rate of 99.99%

Some Database Vendors

- **OCLC**
  - began as bibliographic utility for cooperative cataloguing of library materials
  - now offers access to over 80 databases
  - full text and images from thousands of journals online
- **H.W. Wilson Company**
  - began producing print indexes in 1898
  - now offers 70+ databases to the public, school and college library market
  - both produces databases and provides access to them

IR from Document Databases

- For historical reasons, retrieval based on Boolean search
- While Boolean retrieval still prevalent, trend towards ranked retrieval



IR from Document Databases

- Typical Boolean search

```
BEGIN 4
File    4:INSPEC 1983-2006/Jun W4
        (c) 2006 Institution of Electrical Engineers

      Set   Items   Description
      ---   -----   -----------
?

S WWW OR WEB
            8819   WWW
           59907   WEB
      S1   63381   WWW OR WEB
?

S INFORMATION()(FORAGING OR SCENT)
          566959   INFORMATION
             435   FORAGING
              79   SCENT
      S2      35   INFORMATION()(FORAGING OR SCENT)
?

S S1 AND S2
```

IR from Document Databases

- TARGET requires users to eliminate terms that are not useful for the search

```
? b55
File  55:Biosis Previews(R) 1993-2007/Mar W2 (c) 2007 The Thomson Cor

? target
Input search terms separated by spaces (e.g., DOG CAT FOOD). You can
enhance your TARGET search with the following options:
  - PHRASES are enclosed in single quotes (e.g., 'DOG FOOD')
  - SYNONYMS are enclosed in parentheses (e.g., (DOG CANINE))
  - SPELLING variations are indicated with a ? (e.g., DOG? to search
  - Terms that MUST be present are flagged with an asterisk (e.g., DO

Q = QUIT   H = HELP

? komodo dragon food diet nutrition

Your search will retrieve up to 50 of the statistically most relevant
Searching 2006-2007 records only ...Processing Complete
Your search retrieved 50 records.
```

IR in Organizations

- **Enterprise Search:** search within any organization with digital textual materials, including
  - external Web site
  - company intranet
  - email, database records and shared documents
- Distinct from Web search in various aspects
  - users are employees with specific information needs
  - link structure within documents is limited
  - users do not care whether a document is popular
  - content is usually reliable
- Enterprise search is discussed in great detail in Chapter 15