Web-scale discovery indexes

Eric Lease Morgan
University of Notre Dame

August 18, 2009

Outline

- Introduction
- The problems to be solved
- What versus how; two case studies
- Indexes, not databases
- Next steps; a do-it-yourself recipe
- Plan B
- Do-it-yourself and Plan B compared
- Web-scale indexes and “next generation” library catalogs

Problems to be solved

What is the purpose of our libraries? Why do they exist? What are the problems they are trying to solve?

The answers are not definitive, especially when applied to individual libraries…


© infomotions.com


MACKINAC ISLAND PUBLIC LIBRARY
World
© infomotions.com


Geisel Library
© infomotions.com

The whats of librarianship

Libraries collect, preserve, organize, and disseminate data, information, and knowledge for the purposes of making the work of their respective communities easier.

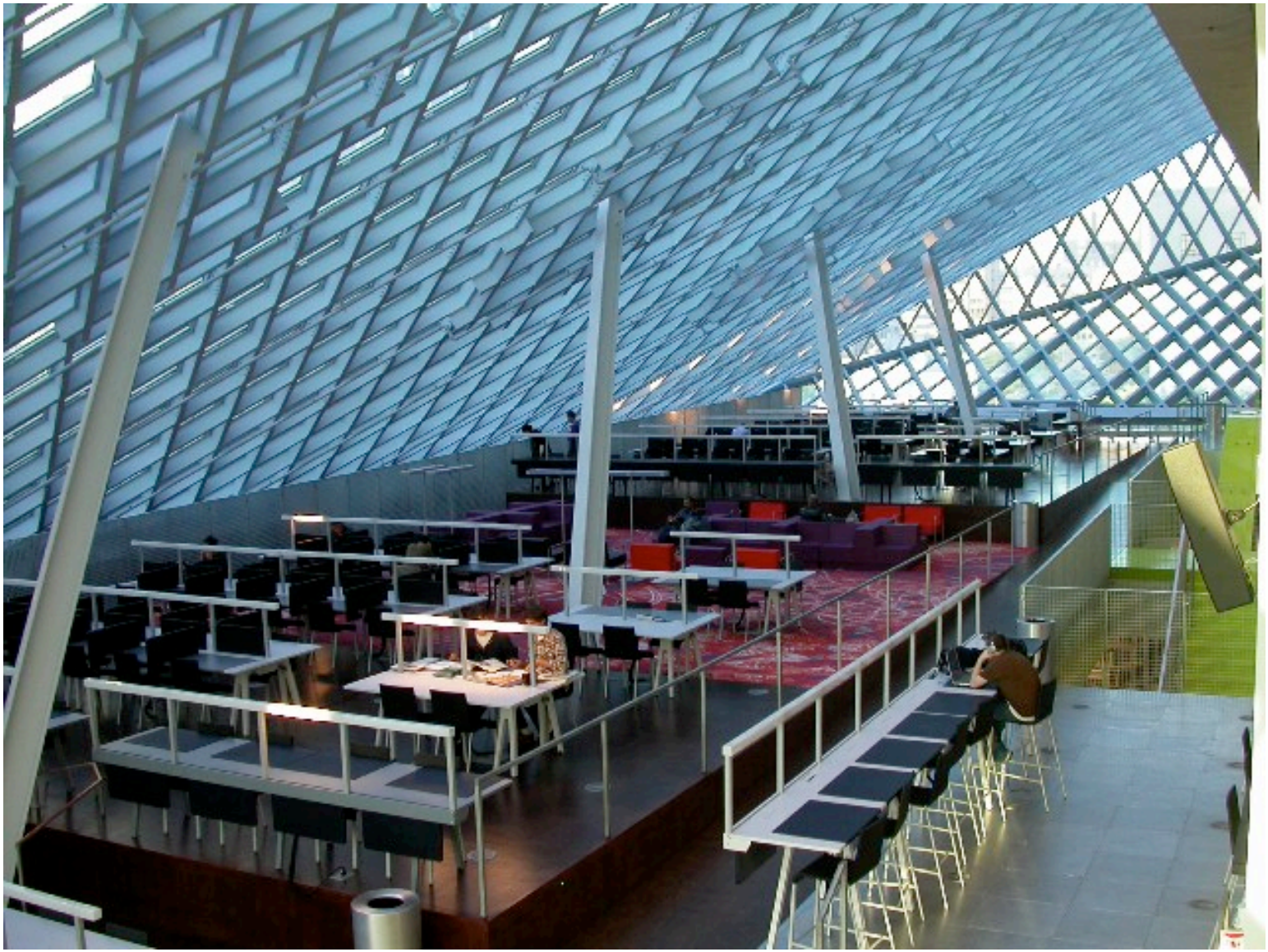
To one degree or another, just about everything us librarians do can be associated with one of these processes.

These things are the whats of librarianship, and they *change very slowly*.

THE WORLD'S RELIGIONS
SMART
Romeo and Juliet
SHAKESPEARE
A Midsummer Night's Dream
SHAKESPEARE
Where the Sidewalk Ends
SILVERSTEIN
© infomotions.com


© infomotions.com


© infomotions.com



The hows of librarianship

The hows of librarianship are the things of our everyday work, our day-to-day operations, the specific workflows within each of our libraries.

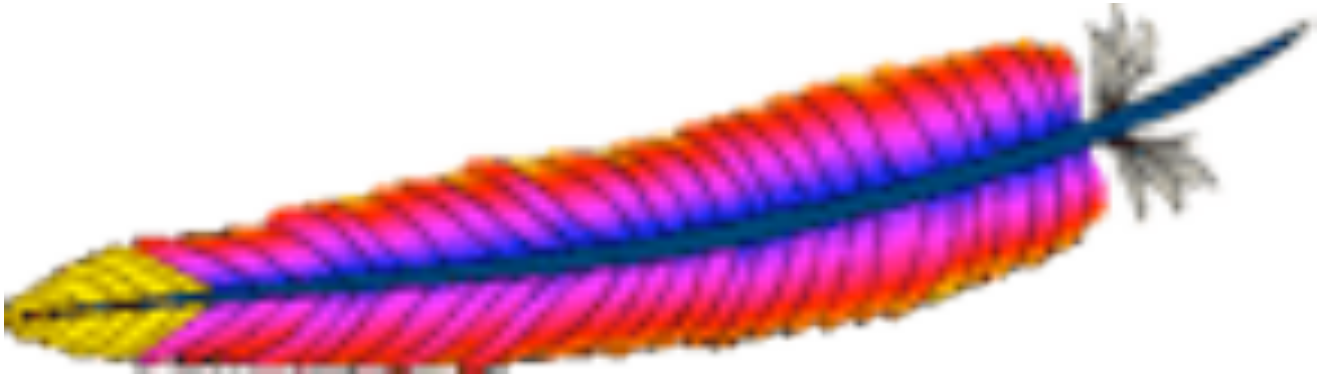
The hows of librarianship change at a much faster pace, and these *changes are usually driven by technology*.

Catalogs are a good example

Journal indexes are another

Indexes, not databases

With the advent of freely available, industrial strength indexers – not databases – we are beginning to see another evolutionary step in the development of the library catalog and journal article index.



Lucene


INDEXI
NDEXIN
DEXIND
EXINDE
X DATA

Powered by

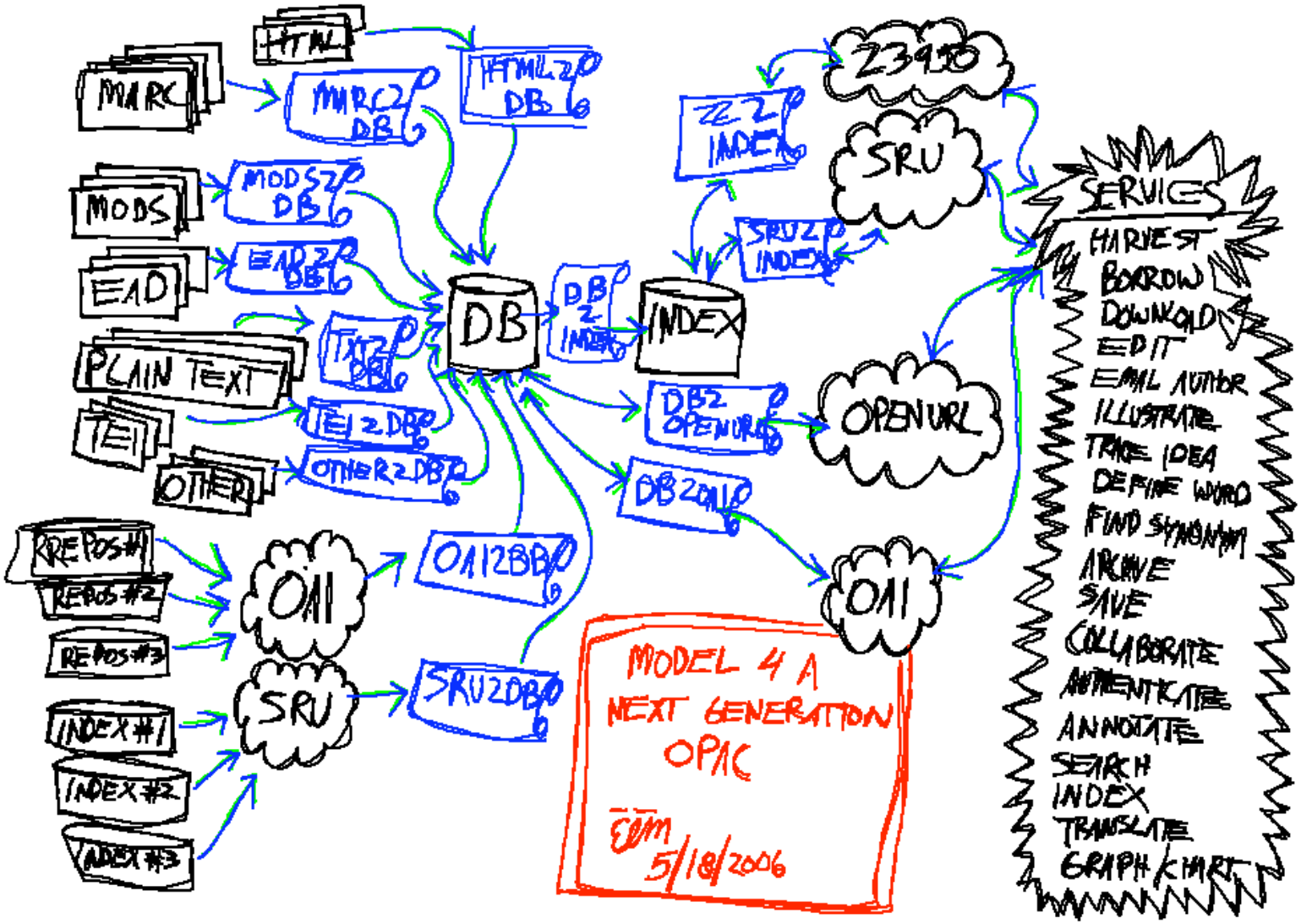
Swish-e

“Smart” computer indexes

```
# calculate term frequency/inverse document frequency
sub tfidf {

  my $n = shift;  # number of times found in document
  my $t = shift;  # total number of words in document
  my $d = shift;  # total number of documents
  my $h = shift;  # number of hits in the corpus

  my $tfidf = 0;

  if ( $d == $h ) { $tfidf = ( $n / $t ) }
  else { $tfidf = ( $n / $t ) * log( $d / $h ) }

  return $tfidf;

}
```

Next steps

What process can be used to take advantage of this environment? What are the next steps? The creation and maintenance of a combined book/journal article index is an excellent example…


HTML
MARC
MARC2 DB
HTML2 DB
MODS
MODS2 DB
EAD
EAD2 DB
PLAIN TEXT
TXT2 DB
TEI
TEI 2 DB
OTHER
OTHER2DB
DB
DB 2 INDEX
INDEX
Z39.50
Z2 INDEX
SRU
SRU2 INDEX
SERVICES
HARVEST
BORROW
DOWNLOAD
EDIT
EMAIL AUTHOR
ILLUSTRATE
TRADE IDEA
DEFINE WORD
FIND SYNONYM
ARCHIVE
SAVE
COLLABORATE
AUTHENTICATE
ANNOTATE
SEARCH INDEX
TRANSLATE
GRAPH/CHART
DB2 OPENURL
OPENURL
DB2OAI
OAI
OAI2DB
REPOS#1
REPOS#2
REPOS#3
OAI
SRU
SRU2DB
INDEX#1
INDEX#2
INDEX#3
MODEL 4 A
NEXT GENERATION
OPAC
EJM
5/18/2006

A recipe

1. Allocate resources
2. Charge the group
3. Wait three months
   - ❑ Dump MARC records
   - ❑ Export metadata from repositories
   - ❑ Harvest and/or mirror article and etext content and metadata
   - ❑ Feed all metadata to indexer
   - ❑ Design simple user interface to search the index
4. Ask for an update
5. Go to Step #3 four times
6. Evaluate
7. Share your experience
8. If the process was successful, then go to Step #1
9. Otherwise, consider Plan B


© infomotions.com


Big Breakfast
KIDS CORNER
© infomotions.com


© infomotions.com

Plan B

For any number of reasons the do-it-yourself approach is not feasible for you and your library…


© infomotions.com


© infomotions.com


© infomotions.com


© infomotions.com

Plan B

…If this is the case, then you might have someone else do the work for you, and they will:

1. Decide what content to include (collections)
2. Collect it (acquisitions)
3. Normalize it (cataloging)
4. Index it (systems)
5. Provide access to the index (public service)

Do-it-yourself and Plan B compared

Neither the do-it-yourself nor the Plan B approach are perfect. There are strengths, opportunities, weaknesses, and threats in both.


9
© infomotions.com


© infomotions.com


© infomotions.com


© infomotions.com

Web-scale indexes – the right direction

But search is not enough

Search and access are not really the problems to be solved. Instead we need to be devising ways to make content more useful and placing it into context: annotate, apply concordance, compare & contrast, create flip book, create tag cloud, graph, hilight, plot on map, rate, review, summarize, trace author and citation, translate, etc.

Thank you

http://www.library.nd.edu/daiad/morgan/musings/web-scale/