Evidence from Metadata

LBSC 796/INFM 718R

Session 9: April 6, 2011

Douglas W. Oard

Problems with “Free Text” Search

- Homonymy
  - Terms may have many <u>unrelated</u> meanings
  - Polysemy (related meanings) is less of a problem
- Synonymy
  - Many ways of saying (nearly) the same thing
- Anaphora
  - Alternate ways of <u>referring to</u> the same thing

Behavior Helps, But not Enough

- Privacy limits access to observations
- Queries based on behavior are hard to craft
  - Explicit queries are rarely used
  - Query by example requires behavior history
- “Cold start” problem limits applicability

A “Solution:” Concept Retrieval

- Develop a concept inventory
  - Uniquely identify concepts using “descriptors”
  - Concept labels form a “controlled vocabulary”
  - Organize concepts using a “thesaurus”
- Assign concept descriptors to documents
  - Known as “indexing”
- Craft queries using the controlled vocabulary

Top: Computers: Software: Information Retrieval *(104)*

Description

- Classification@ *(14)*
- Data Clustering@ *(215)*
- Fulltext *(28)*
- GILS *(3)*
- Internet Search Engines@ *(313)*
- Ranking *(39)*
- References *(1)*
- Text Clustering@ *(24)*
- Visual Information *(6)*
- Web Clustering *(9)*

See also:

- Computers: Software: File Management: Search *(46)*
- Computers: Software: Internet: Servers: Search *(55)*
- Reference: Knowledge Management: Knowledge Retrieval *(40)*
- Reference: Libraries: Library and Information Science: Software *(133)*

This category in other languages:

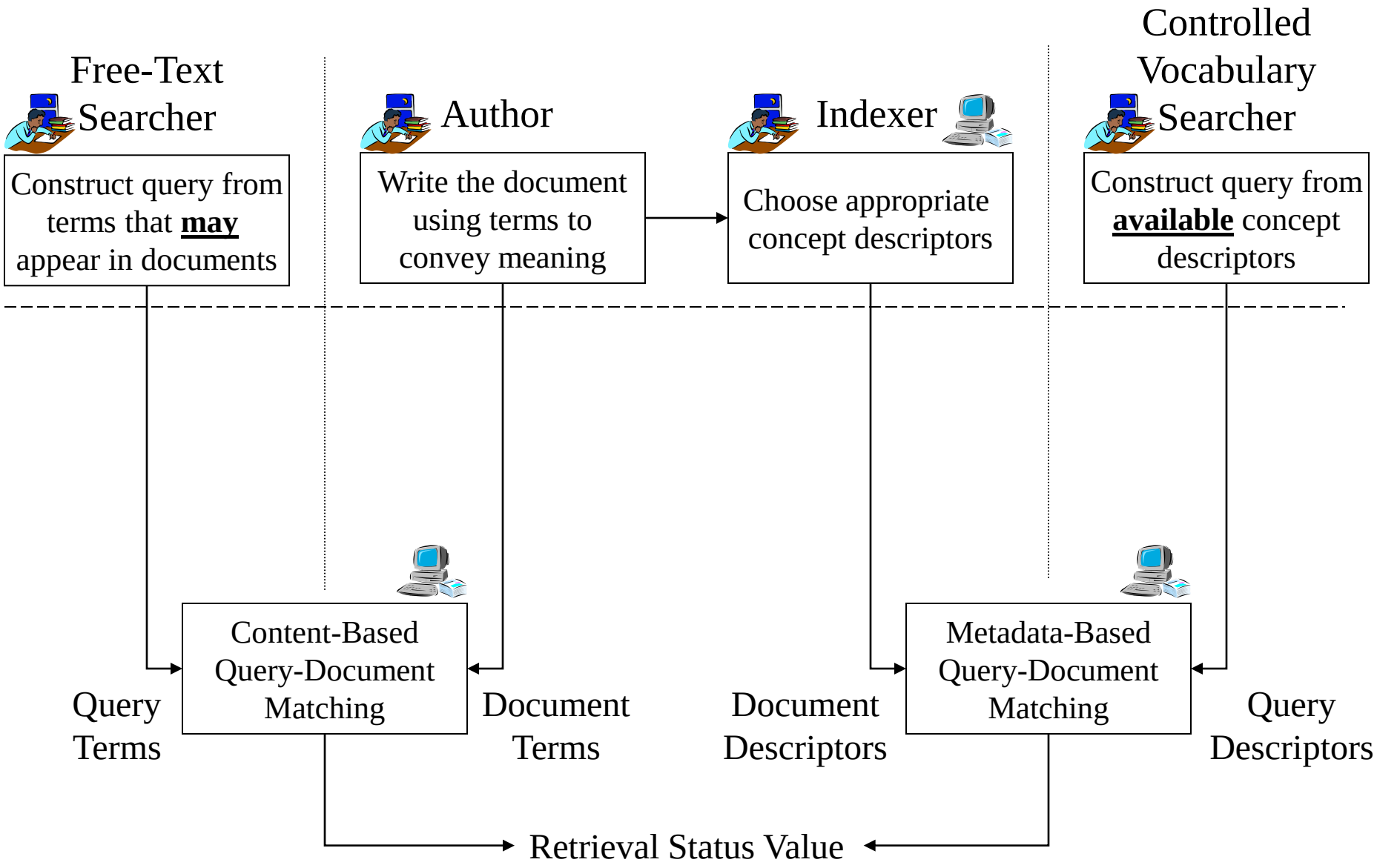
Dutch *(73)*

- AgentWeb: Information Retrieval and Knowledge Management - IR and KM resources specifically relating to intelligent software agents. Includes a wide variety of web resources with good descriptions.
- The Center for Intelligent Information Retrieval - University of Massachusetts research lab focused on efficient access to large, heterogeneous, distributed, text and multimedia databases.
- Clairvoyance Corporation - Develops a suite of component technologies for unstructured text management and analysis. Features overviews of technologies and research initiatives, with company background.
- Collexis - A global company developing software for knowledge retrieval. Collexis both retrieves data and discovers relationships between items via clustering and/or aggregation.
- Delphes Technologies International - Publisher of information retrieval software for personal and corporate knowledge management using natural language technology under the brand name Diogene.
- Extensio - An information integration solution. It makes information from ERP implementations, CRM databases, custom applications, EAI and EIP solutions and the Internet, available on request.
- The Glasgow Information Retrieval Group - Has a research program aimed at giving better access to multi-media information.
- Index Data - Offers courses, software solutions, consultancy aid, and support, relating to Z39.50, Dublin Core, Metadata, and XML.
- Information Retrieval - An online book by C. J. van Rijsbergen, University of Glasgow.
- Information Retrieval Research - An up-to-date overview of research in the field of information retrieval.
- Javaisis 3.0 - JavaIsis is an open source Java application by which you can manage a CDS/ISIS database with any Java Virtual Machine.
- Knowledge Navigation Suite - A suite of information indexing and classification tools that supports information sharing and textual data mining based on natural language processing, statistical pattern analysis, and neural networks techniques. Supports large-scale terabyte data analysis and visualization.
- Modern Information Retrieval - A recent IR book, covering algorithms, implementation, query languages, user interfaces, and multimedia and web retrieval.
- MultiCentrix - Software for information mapping, knowledge management, and computer aided thinking.
- Resources for Text, Speech and Language Processing - A collection of resources in a variety of fields related to text, speech and language processing. These include computational linguistics, information retrieval and machine learning. Here you can find pointers to useful Web sites, as well as lists of relevant books, newsgroups and mailing lists.
- Search-Science - Computer scientists writes about topics usually related to information retrieval (i.e. search results).
- Text REtrieval Conference (TREC) - An annual information retrieval conference and competition, the purpose of which is to support and further research within the information retrieval community.
- Willow - A now discontinued Z39.50 bibliographic information retrieval tool from University of Washington.

- Usenet comp.theory.info-retrieval - news: - Google Groups
- Usenet comp.infosystems.search - news: - Google Groups

- "**Information Retrieval**" search on: AltaVista - A9 - AOL - Ask - Clusty - Gigablast - Google - Lycos - MSN - WiseNut - Yahoo

Two Ways of Searching



Boolean Search Example

Document 1

The quick brown fox jumped over the lazy dog's back.
[Canine] [Fox]

Document 2

Now is the time for all good men to come to the aid of their party.
[Political action] [Volunteerism]

Descriptor	Doc 1	Doc 2
Canine	0	1
Fox	0	1
Political action	1	0
Volunteerism	1	0

- Canine AND Fox
  - Doc 1
- Canine AND Political action
  - Empty
- Canine OR Political action
  - Doc 1, Doc 2

Applications

- When implied concepts must be captured
  - Political action, volunteerism, …
- When terminology selection is impractical
  - Searching foreign language materials
- When no words are present
  - Photos w/o captions, videos w/o transcripts, …
- When user needs are easily anticipated
  - Weather reports, yellow pages, …

Agenda

- Designing metadata
- Generating metadata
- Semantic Web
- Putting the pieces together

Aspects of Metadata

- What kinds of objects can we describe?
  - MARC, Dublin Core, FRBR, …
- How can we convey it?
  - MODS, RDF, OAI-PMH, METS
- What can we say?
  - LCSH, MeSH, PREMIS, …
- What can we do with it?
  - Discovery, description, reasoning

Functional Requirements for Bibliographic Records (FRBR)

- Work (e.g., a specific play)
  - Expression (e.g., a specific performance)
    - Manifestation (e.g., a specific publisher's DVD)
      - Item (e.g., a specific DVD)
- Responsible Entities (person, corporate body)
- Subject (concept, object, event, place)

FRBR in OCLC's FictionFinder

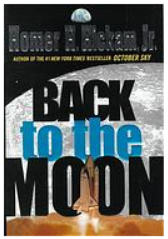
FictionFinder OCLC Research *beta*

Project Page | Feedback | Known Problems | Exit

Browse | Search GO [Advanced]

You searched: Basic Index for **nasa**

Back to Results << Previous Work **5** of **43** Next > Find Any edition



Back to the moon : a novel
Hickam, Homer H., 1943-

6 editions, in 2 languages, held by 1324 libraries

Summary: A cosmic romp in which Jack Medaris, a former astronaut who is a widower, saves the female crew of a NASA spaceship from an attack by chauvinist astronauts. The ladies take him to the moon, he finds romance and brings back a rare helium element which will help his business.

Genres: Adventure fiction | Adventure stories | Science fiction

Settings: Moon

Subjects: Lunar exploration | Helium —Isotopes [+]

Audience: Kids General Special

Editions | Genres | Characters | Settings

Narrow by Languages: All (6)

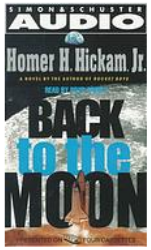
	Title / Author	OCLC #	Date	Langu
1.	Back to the moon : Homer H. Hickam, Jr	40979898	1999	Englis
2.	Back to the moon Homer H. Hickam	41713450	1999	Englis
3.	Back to the moon : Homer H. Hickam, Jr	43890225	2000	Englis
4.	Back to the moon Homer H. Hickam	42765643	1999	Englis
5.	Ben yue zhui qi ling / Xi kan mu zhu ; Wu Hong yi = Back to the moon / by Homer H. Hickam, Jr	47716097	2000	Chine
6.	Back to the moon Homer H. Hickam, Jr	49832301	1999	Englis

FictionFinder OCLC Research *beta*

Project Page | Feedback | Known Problems | Exit

Browse | Search GO [Advanced]

Back to Work < Previous Edition 4 of 6 Next > Find This Edition



Back To the Moon A Novel
Hickam, Homer H., 1943- / Homer H. Hickam. [sound recording] :

Edition: Library ed.

Date: 1999.

Language: English

Publisher: Prince Frederick, MD : Recorded Books, 1999.

ISBN: 0671046403

OCLC: 42765643

Citations | Details | Excerpt | Reviews | Table of Contents

Details

Summary: Jack Medaris, a man of science driven by the memory of the woman who once inspired him, risks everything to sidetrack the space shuttle Columbia and take it on an unscheduled detour to the moon. When the meticulously plotted launch goes fatally wrong, and payload specialist Penny High Eagle further complicates Jack's plan, he must confront unforeseen challenges both in space and on the ground.

Settings: Moon

Performer: Read by Boyd Gaines.

Dublin Core

- Goals:
  - Easily understood, implemented and used
  - Broadly applicable to many applications
- Approach:
  - Intersect several standards (e.g., MARC)
  - Suggest only “best practices” for element content
- Implementation:
  - Initially 15 optional and repeatable “elements”
    - Refined using a growing set of “qualifiers”
  - Now extended to 22 elements

Dublin Core Elements (version 1.1)

Content

- Title
- Subject [LCSH, MeSH, …]
- Description
- Type
- Coverage [spatial, temporal, …]
- Related resource
- Rights

Instantiation

- Date [Created, Modified, Copyright, …]
- Format
- Language
- Identifier [URI, Citation, …]

Responsibility

- Creator
- Contributor
- Source
- Publisher

Resource Description Framework

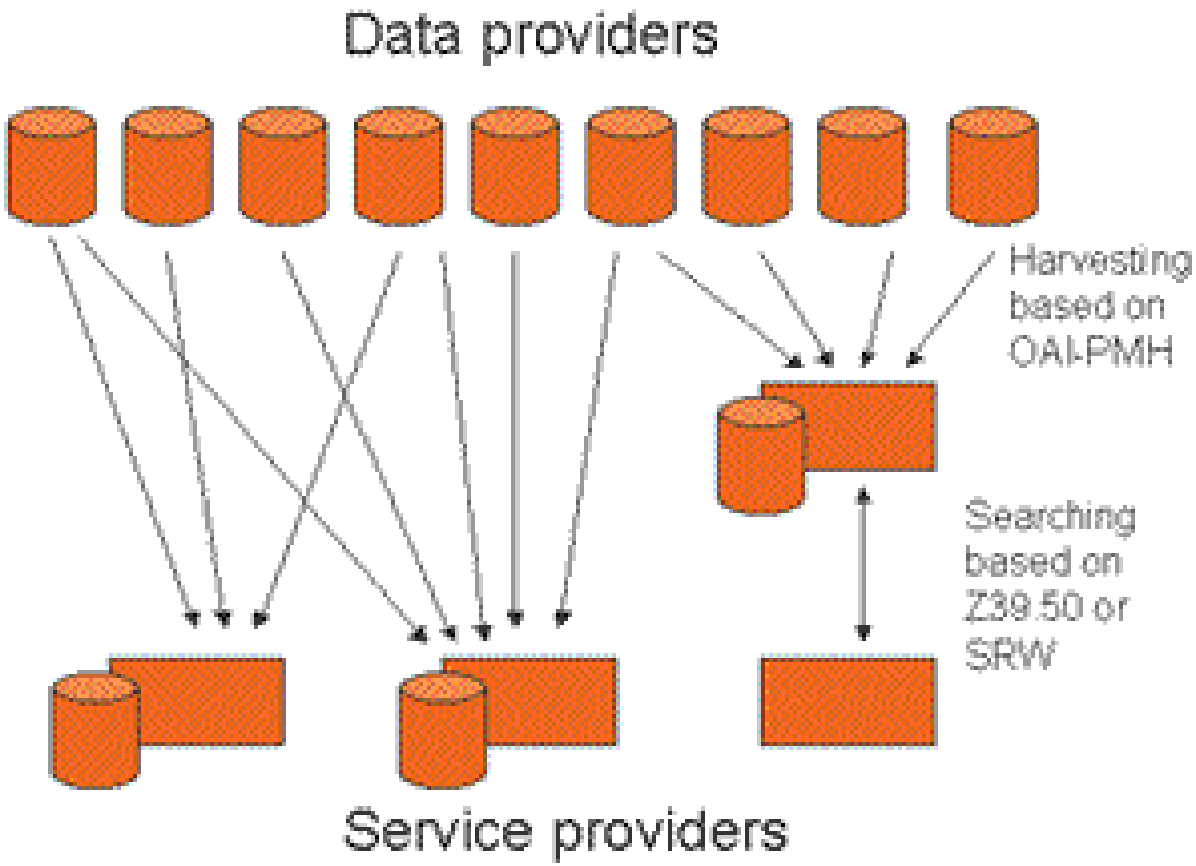
- XML schema for describing resources
- Can integrate multiple metadata standards
  - Dublin Core, P3P, PICS, vCARD, …
- Dublin Core provides a XML “namespace”
  - DC Elements are XML “properties
    - DC Refinements are RDF “subproperties”
  - Values are XML “content”

A Rose By Any Other Name …

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">

  <rdf:Description
    rdf:about="http://media.example.com/audio/guide.ra">
    <dc:creator>Rose Bush</dc:creator>
    <dc:title>A Guide to Growing Roses</dc:title>
    <dc:description>Describes process for planting and nurturing
                    different kinds of rose bushes.</dc:description>
    <dc:date>2001-01-20</dc:date>
  </rdf:Description>
</rdf:RDF>
```

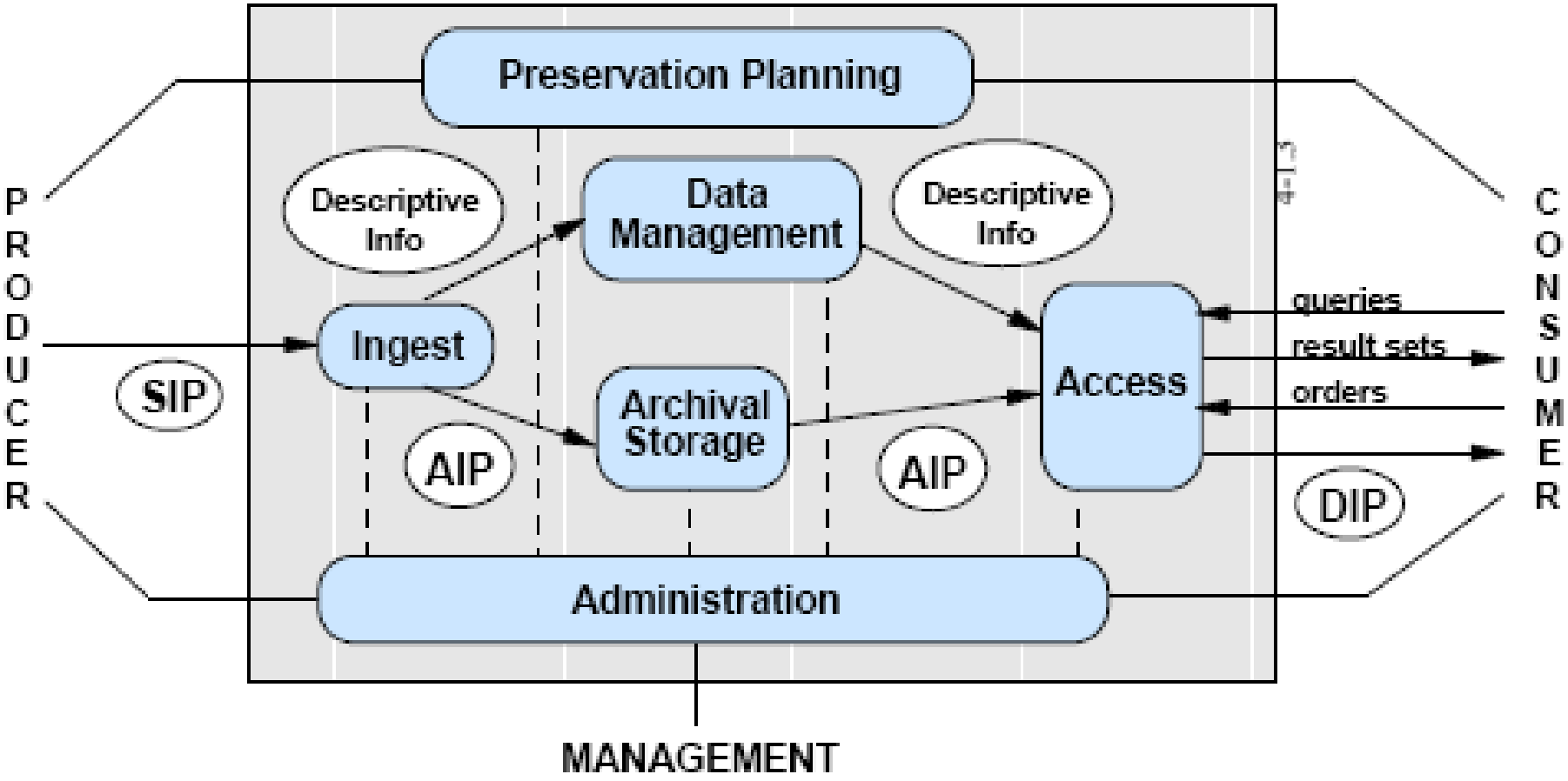
Open Archives Initiative-Protocol for Metadata Harvesting (OAI-PMH)



Metadata Encoding and Transmission Standard (METS)

- Descriptive metadata (e.g., subject, author)
- Administrative metadata (e.g., rights, provenance)
- Technical metadata (e.g., resolution, color space)
- Behavior (which program can render this?)
- Structural map (e.g., page order)
  - Structural links (e.g., Web site navigation links)
- Files (the raw data)
- Root (meta-metadata!)

Open Archival Information System (OAIS) Reference Model



Agenda

- Designing metadata
- ➢ Generating metadata
- Semantic Web
- Putting the pieces together

Thesaurus Design

- Thesaurus must match the document collection
  - Literary warrant
- Thesaurus must match the information needs
  - User-centered indexing
- Thesaurus can help to guide the searcher
  - Broader term (“is-a”), narrower term, used for, …

Challenges

- Changing concept inventories
  - Literary warrant and user needs are hard to predict
- Accurate concept indexing is expensive
  - Machines are inaccurate, humans are inconsistent
- Users and indexers may think differently
  - Diverse user populations add to the complexity
- Using thesauri effectively requires training
  - Meta-knowledge and thesaurus-specific expertise

Machine-Assisted Indexing

- Goal: Automatically suggest descriptors
  - Better consistency with lower cost
- Approach: Rule-based expert system
  - Design thesaurus by hand in the usual way
  - Design an expert system to process text
    - String matching, proximity operators, …
  - Write rules for each thesaurus/collection/language
  - Try it out and fine tune the rules by hand

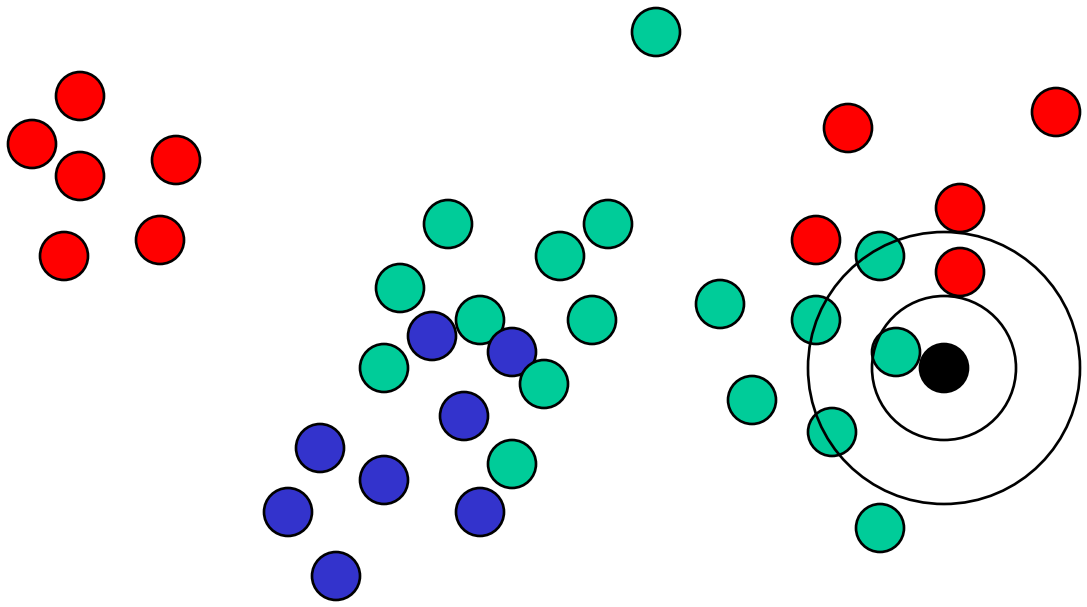
Machine-Assisted Indexing Example

Access Innovations system:

```
//TEXT: science
IF (all caps)
  USE research policy
  USE community program
ENDIF
IF (near “Technology” AND with “Development”)
  USE community development
  USE development aid
ENDIF
```

near: within 250 words
with: in the same sentence

Machine Learning: kNN Classifier



“Folksonomies”

delicious from YAHOO!

It's Free! Join Now Sign In

The tastiest bookmarks on the web.
Save your own or see what's fresh now!

Learn More

HIDE INTRO

Search the biggest collection of bookmarks in the universe...

Search Delicious Search

Fresh Bookmarks | Hotlist | Explore Tags

The freshest bookmarks that are flying like hotcakes on Delicious and beyond.

See more recent bookmarks

New bookmarks saved in the last minute 1 2 5

Spy on Your Facebook Competitors with HyperAlerts | social media tools | Social Media Consulting - Convince & Convert 18
SAVE | SHARE
via convinceandconvert.com
facebook tools monitoring alerts socialmedia
11 Related Tweets

ThinkProgress » Bristol Palin's Nonprofit Paid Her Seven Times What It Spent On Actual Teen Pregnancy Prevention 10
SAVE | SHARE
via thinkprogress.org
humour hso sex #guild palin
6 Related Tweets

13 Things You Must Do Every Week As A Startup CEO | betashop SAVE | SHARE 61
via betashop.com
startup business management ceo entrepreneurship
3 Related Tweets

4 Great Firefox Add-Ons For Mac Users SAVE | SHARE 5
via makeuseof.com
firefox add_on apps mac GReader
4 Related Tweets

Microsoft Excel - You asked about VLOOKUP SAVE | SHARE 5
via blogs.office.com

All items tagged **radio** (**create tag description**) → view **popular**

del.icio.us search

« earlier | later »

Playbill Radio save this
by wheelmaker2 to music radio broadway playbill Entertainment ... saved by 12 other people ... 2 mins ago

Rhapsody save this
by srminton to music rhapsody radio streaming entertainment mp3 ... saved by 515 other people ... 3 mins ago

Kasper Hauser's "This American Life" Parody: Episode 1 save this
Sounding like This American Life.
by hansenn to comedy radio thisamericanlife ... saved by 27 other people ... 5 mins ago

Breaking News | Latest News | Current News - FOXNews.com save this
by parcley to radio news ... saved by 2839 other people ... 7 mins ago

Family.org save this
by bastian_balthasar_bux to Family christian Christianity radio news RELIGION reference ... saved by 311 other people ... 16 mins ago

BBC - 1Xtra - Homepage save this
by okajun to reggae radio ... saved by 135 other people ... 17 mins ago

Sound & Spirit save this
by dragonjazz to radio ... saved by 19 other people ... 19 mins ago

http://www.pandora.com/?tc=x-036821-0035-1149 save this
music
by sarah.bierman to radio ... saved by 4 other people ... 20 mins ago

related tags

music
media
audio
scanner
streaming
radiolocator
frequencies
ham
musik
journalism
imported

“Named Entity” Tagging

- Machine learning techniques can find:
  - Location
  - Extent
  - Type

- Two types of features are useful
  - Orthography
    - e.g., Paired or non-initial capitalization
  - Trigger words
    - e.g., Mr., Professor, said, …

File Edit View Go Favorites Help Address D:\RnRDemo.htm

Your query has finished

Rough'n'Ready GTE

Search

Clear

OR

AND

Topic

Person

Organization

Location

Speaker

Text

Story Jewish-Arab relations : Politics and government : Palestinian Arabs : Middle East : Israel : Terroris

5 stories about: Jewish-Arab relations : Politics and government : Palestinian Arabs : Middle East : Israel : Terr

Jewish-Arab relations : Politics and government : Palestinian Arabs : Middle East : Israel : Terrorism : Pale

Jewish-Arab relations : Israel : Middle East : Middle East peace negotiations : Politics and government : P

male 5

Well as all work during president Clinton's trip to New York tonight and he enjoys the performance of the opera Carmen at Lincoln Center and see the scene there is a lot of Broadway. Now earlier today Mr. Clinton announced that the UN united Nations general assembly that he plans to send a nuclear test ban treaty to the Senate the treaty bans all nuclear test explosions and is regarded as a milestone in the arms control. Two israeli security guards were wounded in an early morning shooting in Jordan a government official says three men and a car opened fire on the guard's car wounding both before Skipping guards were treated at a hospital and released it is real several West Bank villages were sealed by israeli soldiers who search for the islamic militants behind two recent suicide bombings in Jerusalem palestinian leader Yasser Arafat says that he believes those was counsel for the bombing case and abroad.

Jewish-Arab relations

Middle East peace negotiations

Middle East

Palestinian self-rule areas

Israel

Politics and government

Arafat, Yasir

Palestinian Arabs

Normalization

- Variant forms of names (“name authority”)
  - Pseudonyms, partial names, citation styles
- Acronyms and abbreviations
- Co-reference resolution
  - References to roles, objects, names
  - Anaphoric pronouns
- Entity Linking

Entity Linking

Example: Bibliographic References

15. Faloutsos, C., Oard, D. (1995). “A Survey of Information Retrieval and Filtering Methods,” avail. As UMIACS-TR-95-33, College Park: U of MD.

[11] Faloutsos, C. and Oard, D. W., "A Survey of Information Retrieval and Filtering Methods", University of Maryland, Technical Report CS-TR-3514, August 1995.

[47] Christos Faloutsos and Douglas W. Oard. A survey of information retrieval and filtering methods. Technical Report CS-TR-3514, University of Maryland, August 1995. http://www.enee.umd.edu/medlab/filter/papers/survey.ps.

[Faloutsos] Christos Faloutsos and Dougtas Oard, A Survey of Information Retrieval and Filtering Method,
<URL:http://www.glue.umd.edu/enee/medlab/filter/papers/survey.ps>

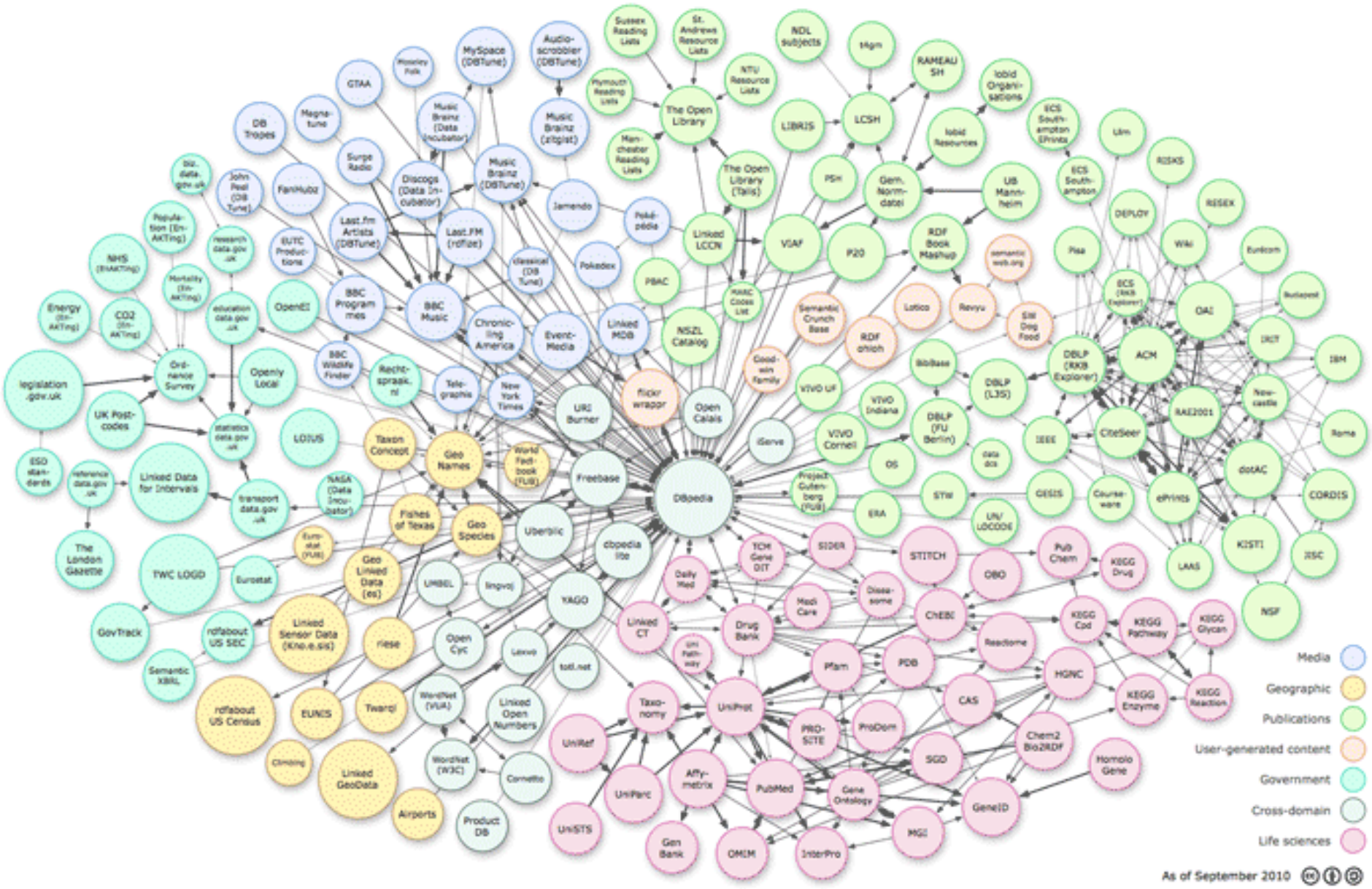
Agenda

- Designing metadata
- Generating metadata
- ➢ Semantic Web
- Putting the pieces together

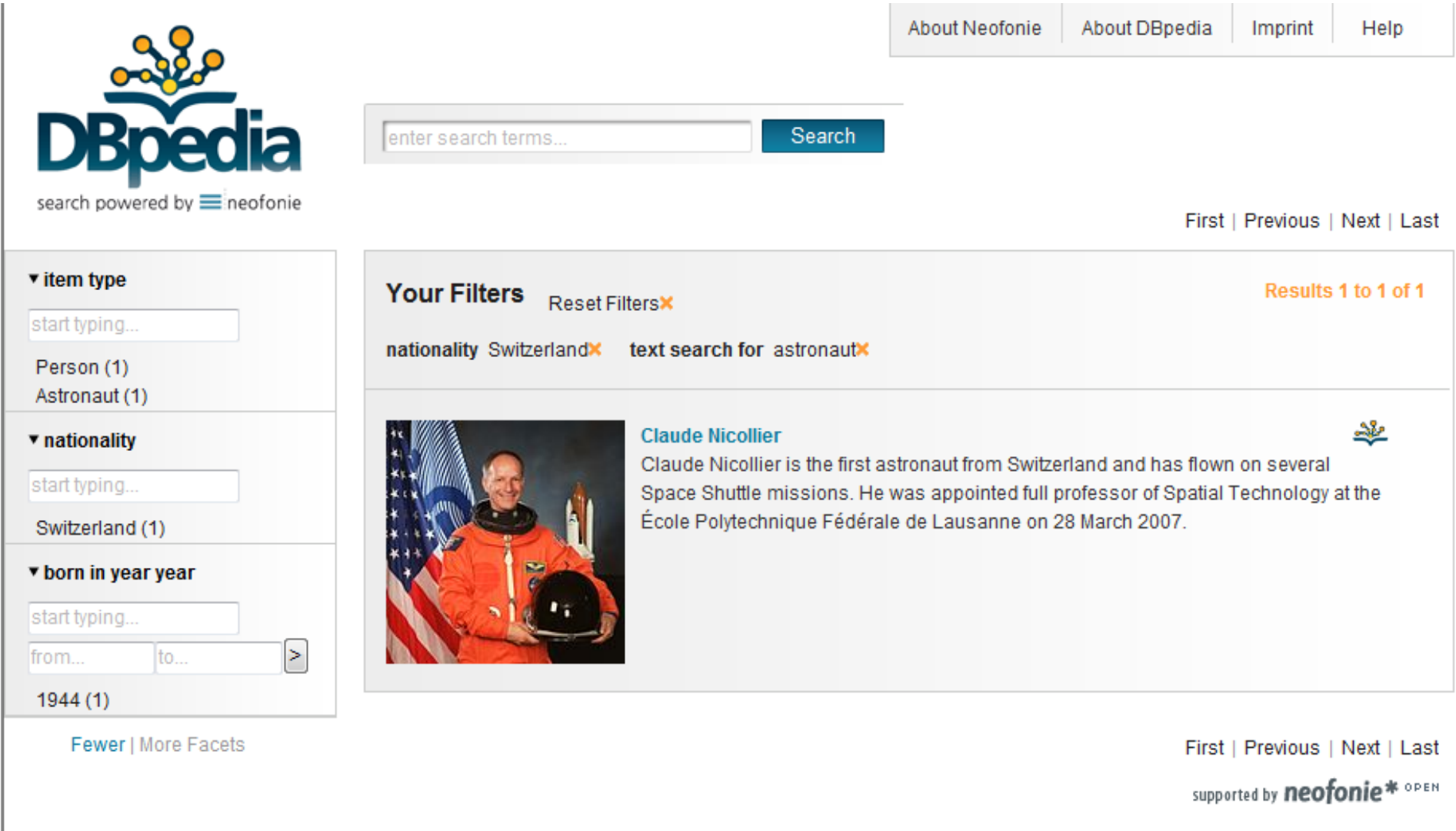
Web Ontology Language (OWL)

```
<owl:Class rdf:about="http://dbpedia.org/ontology/Astronaut">
        <rdfs:label xml:lang="en">astronaut</rdfs:label>
        <rdfs:label xml:lang="de">Astronaut</rdfs:label>
        <rdfs:label xml:lang="fr">astronaute</rdfs:label>
        <rdfs:subClassOf
                rdf:resource="http://dbpedia.org/ontology/Person">
        </rdfs:subClassOf>
</owl:Class>
```

Linked Open Data



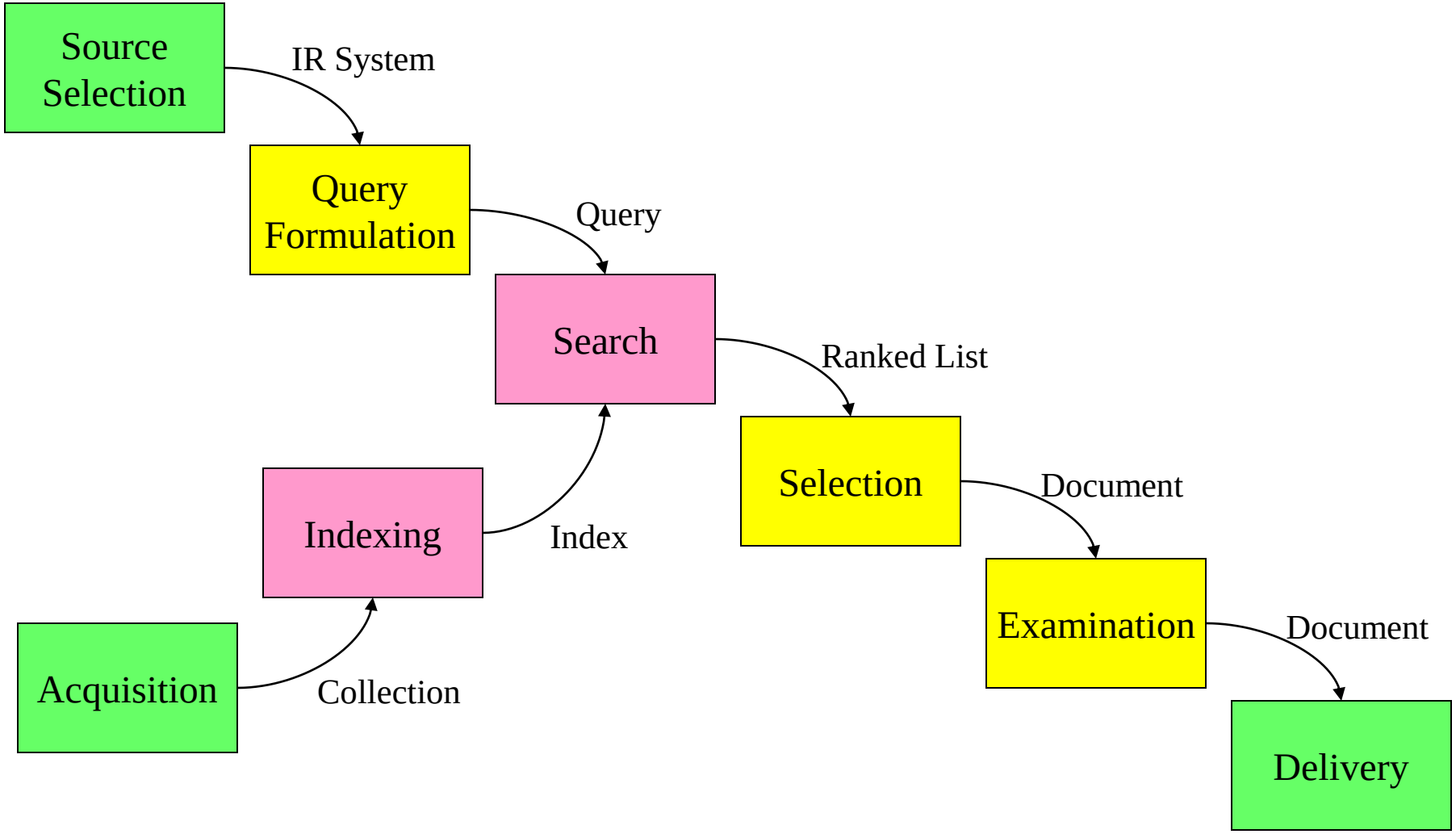
Semantic Web Search



Agenda

- Designing metadata
- Generating metadata
- Semantic Web
- ➢ Putting the pieces together

Supporting the Search Process



Putting It All Together

	Free Text	**Behavior**	**Metadata**
Topicality			
Quality			
Reliability			
Cost			
Flexibility			

Before You Go!

On a sheet of paper, please briefly answer the following question (no names):

What was the muddiest point in today's class?