
Semi-Structured
Unstructured data
Volume
Petabytes
Unstructured content
Data analysis
BIG DATA
Useful
Metadata
AMP
People driven
Decision making
Data
Zettabyte
Framework
Unstructured
Smart content
database
VAST
Concept extraction
Text analytics
Semantic Metadata
Structured

Big Data Content Organization, Discovery, and Management

Marjorie M.K. Hlava
President
Access Innovations, Inc.
Mhlava@accessinn.com

Access Innovations, Inc.. 12/11/2013

Outline

- Big Data
- New Government Initiative
- Content Organization
- Discovery (Search)
- Management
- Skills we bring
- Examples of what we can do

Why do we care about Big Data?

- Data is the new oil – we have to learn how to mine it! Qatar – European Commission Report
- $ 7 trillion economic value in 7 US sectors alone
- $90 B annually in sensitive devices
- An insurance firm with 5 terabytes of data on share drives pays $1.5 m per year
- New McKinsey 4th factor of production
- Land, Labor, Capital, + Data

The Data Deluge – Wired 16.07

Big Data Born

- Google, eBay, LinkedIn, and Facebook were built around Big Data from the beginning.
- No need to reconcile or integrate Big Data with more traditional sources of data and the analytics performed upon them
- No merging Big Data technologies with their traditional IT infrastructures
- Big Data could stand alone, Big Data analytics could be the only focus of analytics
- Big Data technology architectures could be the only architecture.

Integrating Big Data

- Large, well-established orgs
- Must be integrated with everything else that's going on in the company.
- Analytics on Big Data have to coexist with analytics on other types of data.
- Hadoop clusters have to do their work alongside IBM mainframes.
- Data scientists must somehow get along and work jointly with mere quantitative analysts.

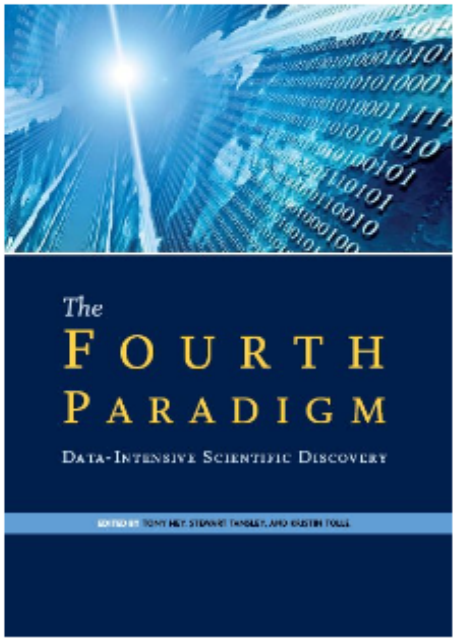
What is Big Data?

- *Big Data is a term applied to data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Big Data sizes are a constantly moving target currently ranging from a few dozen terabytes to many petabytes of data in a single data set. – Wikipedia, May 2011*
- "Unstructured"
- Terabytes, petabytes, zettabytes
- Streaming

New kind of science?

The Fourth Paradigm: Data-Intensive Scientific Discovery

Presenting the first broad look at the rapidly emerging field of data-intensive science



Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets.

The speed at which any given scientific discipline advances will depend on how well its researchers collaborate with one another, and with technologists, in areas of eScience such as databases, workflow management, visualization, and cloud computing technologies.

In *The Fourth Paradigm: Data-Intensive Scientific Discovery*, the collection of essays expands on the vision of pioneering computer scientist Jim Gray for a new, fourth paradigm of discovery based on data-intensive science and offers insights into how it can be fully realized.

Download

- Full text, low resolution (6 ME
- Full text, high resolution (93 M
- By chapter and essay

Purchase from Amazon.com

- Paperback
- Kindle version

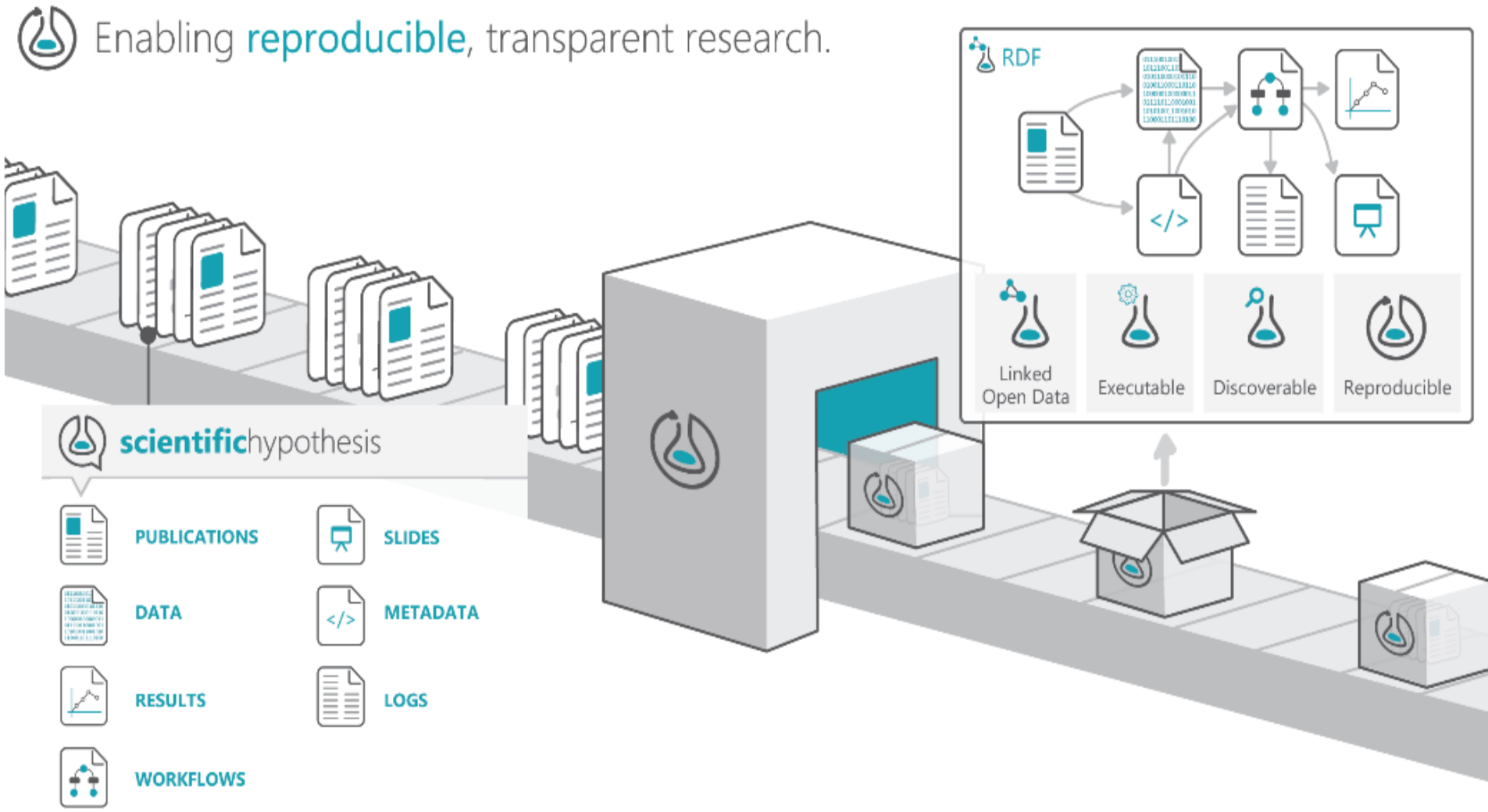
In the news

- Sailing on an Ocean of 0s and
  Magazine)
- A Deluge of Data Shapes a N
  Computing (*New York Times*)

New Special Collections

- Volume, Velocity, Variety
- Ability to deal overwhelmed
- More about methods than data
- Location aware data
- Life streaming
- Insurance claims
- Hubble telescope
- CERN Collections
- Flight data

www.researchobject.org



Unstructured data

- Means untagged or unformatted
- PDF
- Word files
- File shares
- News feeds
- News Data feeds
- Images

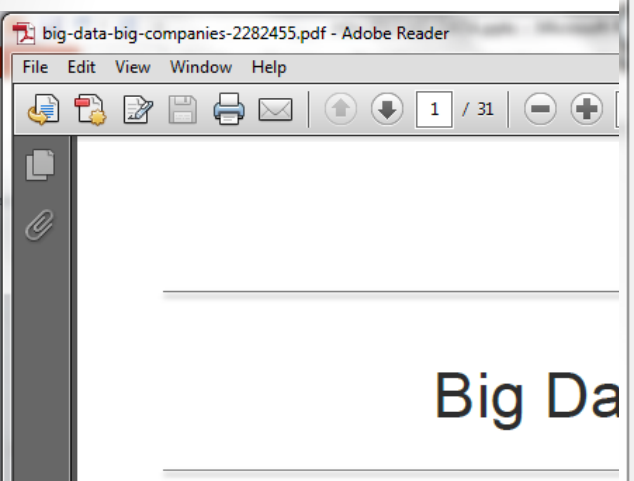
Bit of a misnomer

- All data has some structure and more structure possibilities
- PDF properties
- Word file properties
- File structures

Property tables

- PDF - property tables
- Word files – property tables
- Files shares – implied structure in file names
- News feeds – headers have metadata
- Hubbell telescope feeds
- Images

PDF Prope



Document Properties

Description | Security | Fonts | Custom | Advanced

Description

File:	big-data-big-companies-2282455.pdf
Title:	Big Data in Big Companies
Author:	David DeMots
Subject:	
Keywords:	; null
Created:	5/2/2013 7:44:18 PM
Modified:	5/2/2013 7:44:18 PM
Application:	Microsoft® Word 2013

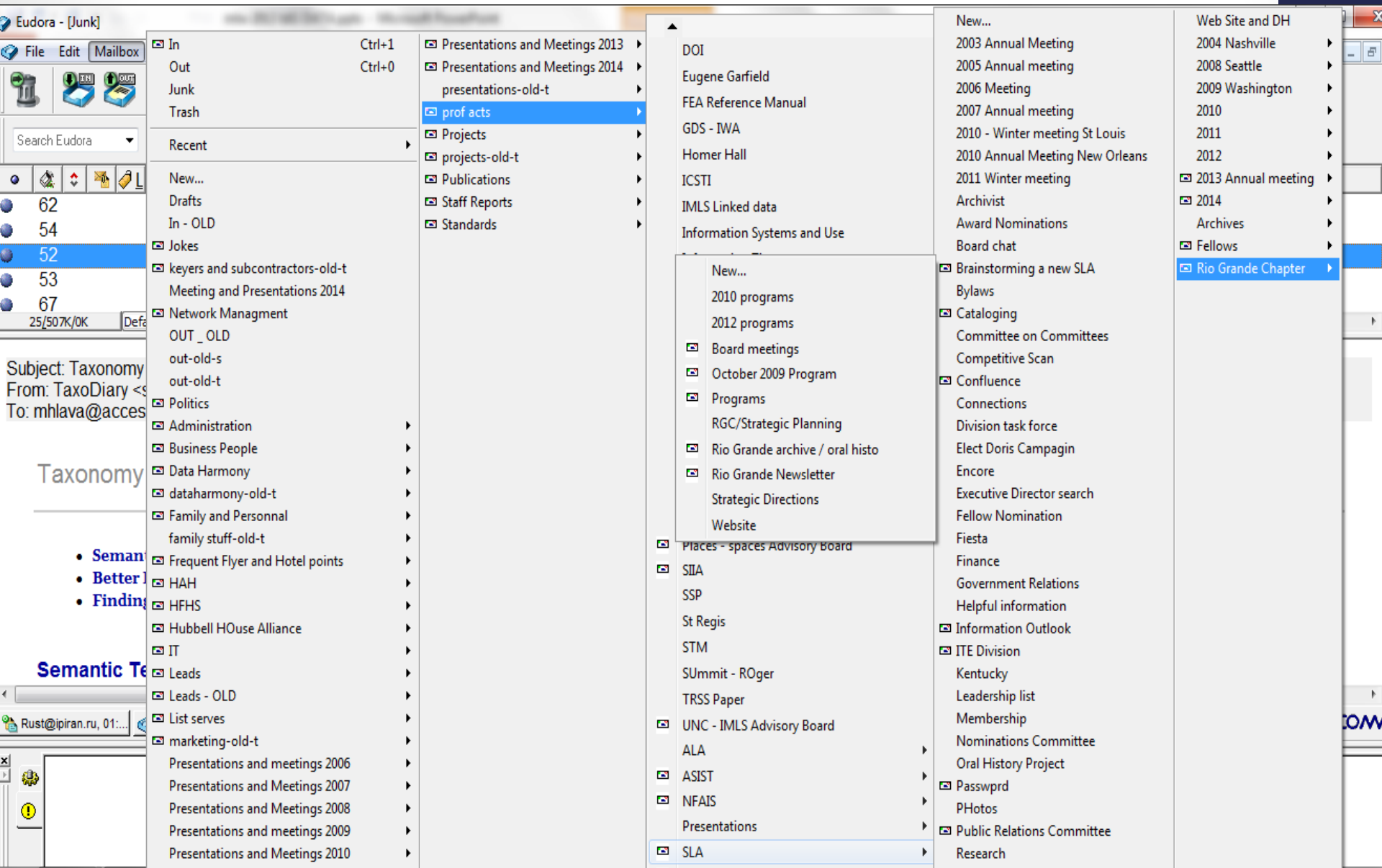
Advanced

PDF Producer:	Microsoft® Word 2013		
PDF Version:	1.5 (Acrobat 6.x)		
Location:	C:\Users\Margie\Documents\		
File Size:	1.78 MB (1,871,197 Bytes)		
Page Size:	8.50 x 11.00 in	Number of Pages:	31
Tagged PDF:	Yes	Fast Web View:	No

OK Cancel

Jill Dyché

File structures



Structu

- XML tagg

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE us-patent-grant SYSTEM "us-patent-grant-v42-2006-08-23.dtd">
<us-patent-grant lang="EN" country="US" dtd-version="v4.2 2006-08-23">
 - <us-bibliographic-data-grant>
    - <publication-reference>
       - <document-id>
            <country>US</country>
            <doc-number>6170828</doc-number>
            <date>20010109</date>
         </document-id>
      </publication-reference>
    - <application-reference appl-type="utility">
       - <document-id>
            <country>US</country>
            <doc-number>09492395</doc-number>
            <date>20000127</date>
         </document-id>
      </application-reference>
      <us-application-series-code>09</us-application-series-code>
    - <classifications-ipcr>
       - <classification-ipcr>
          - <ipc-version-indicator>
               <date>20000101</date>
            </ipc-version-indicator>
            <section>A</section>
            <class>63</class>
            <subclass>F</subclass>
            <main-group>1</main-group>
            <subgroup>00</subgroup>
         </classification-ipcr>
      </classifications-ipcr>
    - <classification-national>
         <country>US</country>
         <main-classification>273292</main-classification>
```

What are the problems?

- Data infrastructure challenges
- “taking diverse and heterogeneous data sets and making them more homogeneous and usable”
- An opportunity?
- All that data – what can it tell us?
- Privacy
- Copyright
- Neurological impact
- Data collection methods

New Government Initiative

The Big Data Senior Steering Group (BDSSG) was formed to identify current Big Data research and development activities across the Federal government, offer opportunities for coordination, and identify what the goal of a national initiative in this area would look like. Subsequently, in March 2012, The White House Big Data R&D Initiative was launched and the BDSSG continues to work in four main areas to facilitate and further the goals of the Initiative.

The National Big Data R&D Initiative

- Fast-growing volume of digital data of digital data
- Advance state-of-the-art core technologies needed to collect, store, preserve, manage, analyze, and share huge quantities of data.
- Harness these technologies to accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning; and
- Expand the workforce needed to develop and use Big Data technologies.

Data to Knowledge to Action

- Advance supporting technologies
  - Big Data
  - Data analytics;
- Educate and expand the Big Data workforce
- Improve key outcomes in economic growth, job creation, education, health, energy, sustainability, public safety, advanced manufacturing, science and engineering, and global development

Data on Data

- January 22, 2013 - Data on Data: Presenting Stakeholder Alignment Data on the Cyberinfrastructure for Earth System Science
- Presentation and discussion with Professor Joel Cutcher-Gershenfeld. Professor Cutcher-Gershenfeld presented information on the NSF EarthCube initiative including stakeholder survey data (approximately 850 responses).

Who is involved?

DARPA

AHRQ

Homeland Security
Science and Technology

NATIONAL ARCHIVES

Office of the National Coordinator for Health Information Technology

U.S. DEPARTMENT OF ENERGY
Office of Science

NNSA
National Nuclear Security Administration

NASA

Groups breakdown

- INTERAGENCY WORKING GROUPS
- Cyber Security and Information Assurance
- High End Computing
- COMMUNITY OF PRACTICE (CoP)
- · Faster Administration of Science and Technology Education and Research
- COORDINATING GROUPS
- Human Computer Interaction and Information Management
- High Confidence Software and Systems
- Large Scale Networking
- Software Design and Productivity
- Social, Economic, and Workforce Implications of IT

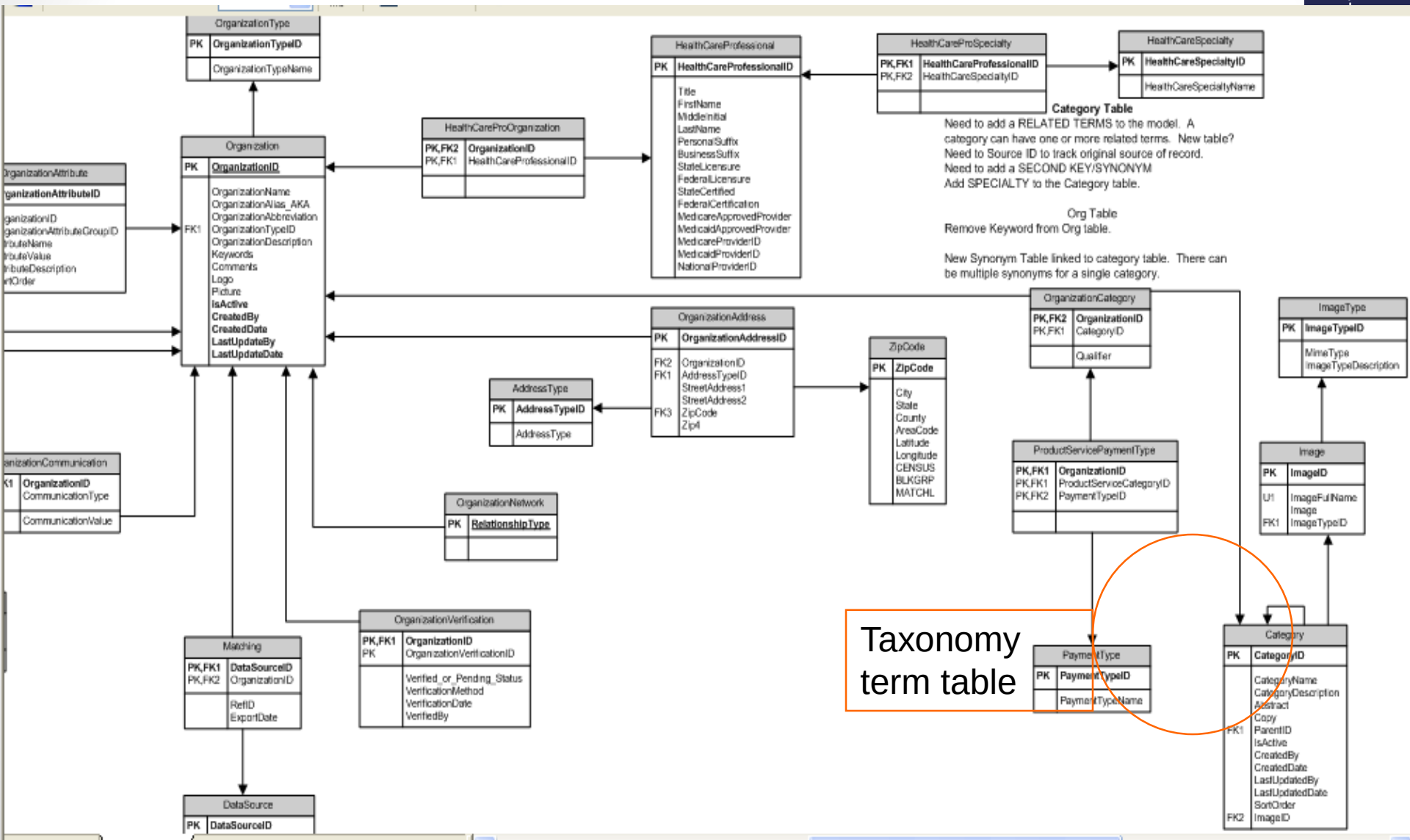
SENIOR STEERING GROUPS (SSGs)

- Big Data
- Cyber Physical Systems
- Cyber Security and Information Assurance Research and Development
- Health Information Technology Research and Development
- Wireless Spectrum Research and Development
- SUBGROUP
- Health Information Technology Innovation and Development Environments Subgroup
- TEAMS
- Joint Engineering Team
- Middleware and Grid Interagency Coordinating Team

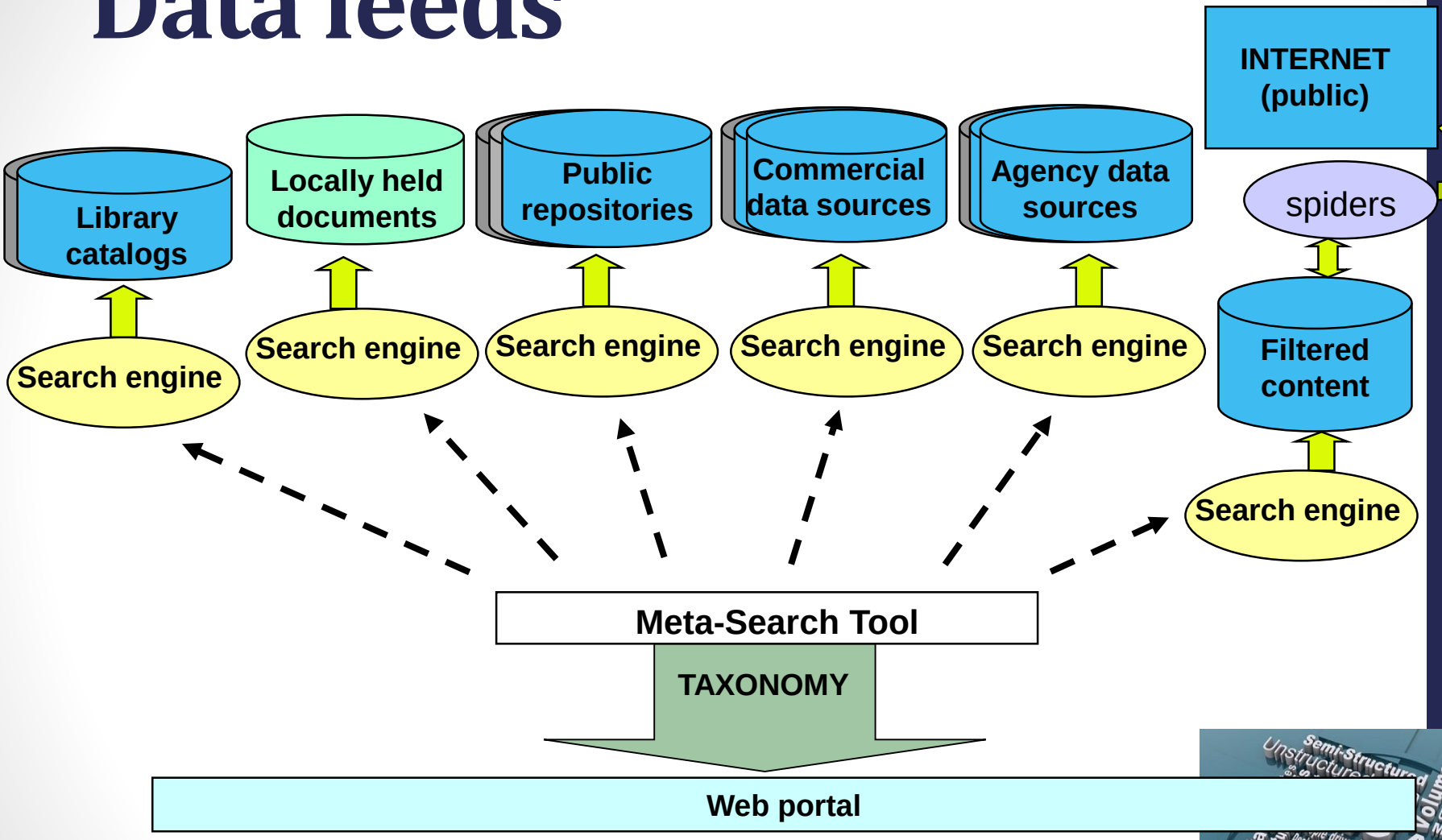
Content Organization

- Data on machines
  - Local
  - Cloud
  - Remote
  - Streaming
- Undifferentiated
- Unstructured
- Needs organization
- Type of database structure
  - RDBMS
  - Object oriented

RDBMS Connection



Data feeds



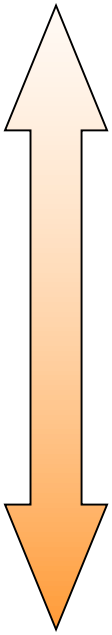
Metadata options

- Structure unstructured data
- Create metadata
- Where to put it?
- Store metadata with the records
  - HTML header
  - Properties tables
  - XML files
- Store metadata in a separate file
  - Database
  - Metadata repository
  - Search system
  - File structure
  - SharePoint application
  - Web interface

Information retrieval starts with a knowledge organization system

- Uncontrolled list
- Name authority file
- Synonym set / ring
- Controlled vocabulary
- *Taxonomy*
- *Thesaurus*
- Ontology
- Topic Map
- Semantic Network

Not complex - $



Highly complex - $$$$

LOTS OF OVERLAP!

Structure of controlled vocabularies

List of words Synonyms Taxonomy Thesaurus

INCREASING COMPLEXITY / RICHNESS

Ambiguity control

Synonym control

Ambiguity control
Synonym control
Hierarchical rel's

Ambiguity cont'l
Synonym cont'l
Hierarchical rel's
Associative rel's

Taxonomy / thesaurus

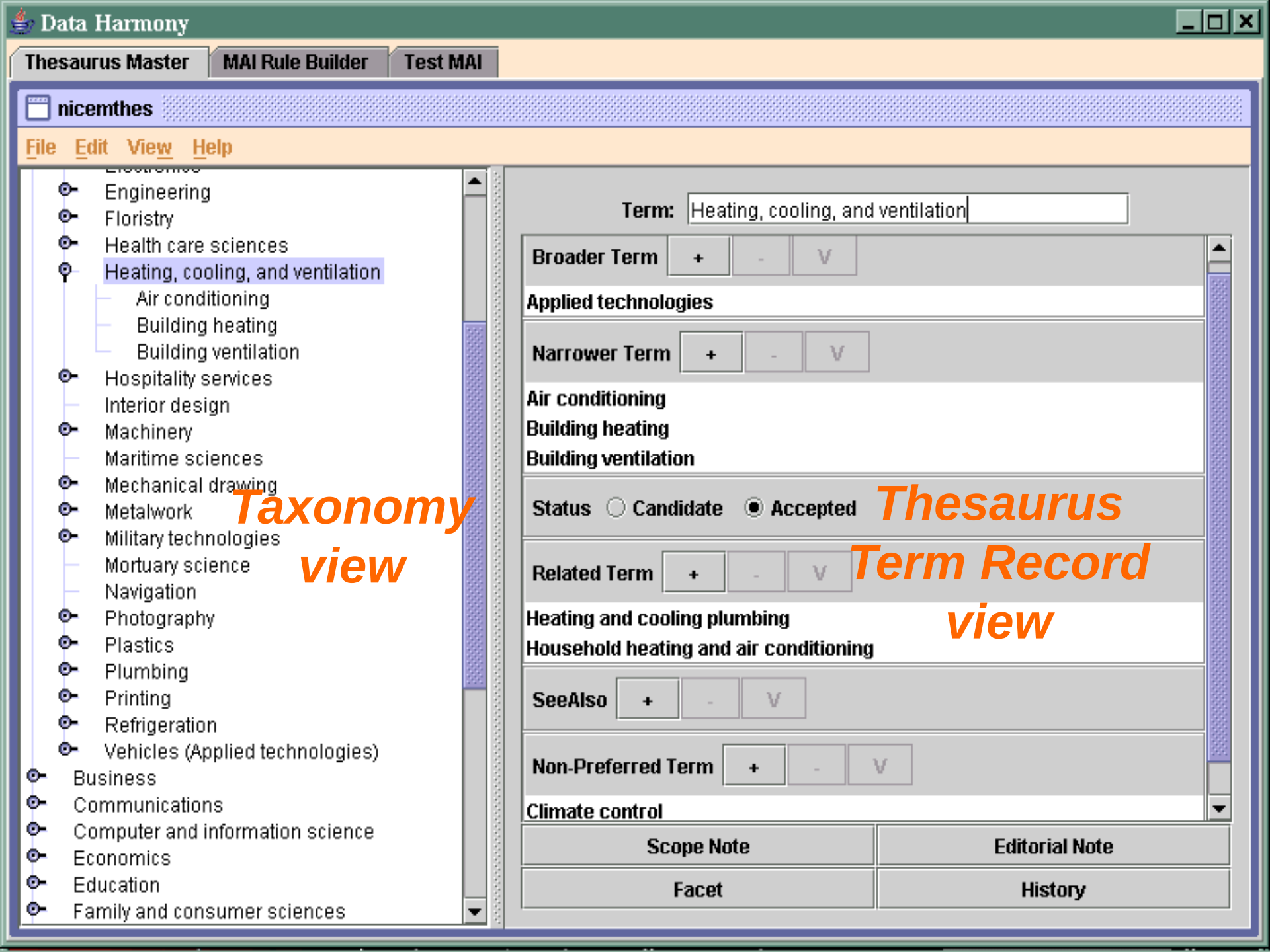
- **Main Term (MT) ***
- **Top Term (TT)**
- **Broader Terms (BT)**
- **Narrower Terms (NT)**
- **Narrower Term Instance**
- Related Terms (RT)
  - See also (SA)
- Non-Preferred Term (NP)
  - Used for (UF), See (S)
- Scope Note (SN)
- History (H)

THESAURUS

TAXONOMY

ONTOLOGY

*a.k.a. subject term, heading, node, index term, keyword, category, descriptor, class


Data Harmony
Thesaurus Master
MAI Rule Builder
Test MAI
nicemthes
File
Edit
View
Help
Engineering
Floristry
Health care sciences
Heating, cooling, and ventilation
Air conditioning
Building heating
Building ventilation
Hospitality services
Interior design
Machinery
Maritime sciences
Mechanical drawing
Metalwork
Military technologies
Mortuary science
Navigation
Photography
Plastics
Plumbing
Printing
Refrigeration
Vehicles (Applied technologies)
Business
Communications
Computer and information science
Economics
Education
Family and consumer sciences
Taxonomy view
Term: Heating, cooling, and ventilation
Broader Term + - V
Applied technologies
Narrower Term + - V
Air conditioning
Building heating
Building ventilation
Status Candidate Accepted
Related Term + - V
Heating and cooling plumbing
Household heating and air conditioning
SeeAlso + - V
Non-Preferred Term + - V
Climate control
Scope Note
Editorial Note
Facet
History
Thesaurus Term Record view

Taxonomies in Context

A taxonomy aspires to be:

- a correlation of the different functional, regional and (possibly) national languages used by a community of practice
- a support mechanism for navigation
- a support tool for search engines and knowledge maps
- an authority for tagging documents and other information objects
- a knowledge base in its own right

Reference: "Taxonomies: the vital tool of information architecture", www.tfpl.com. Courtesy of Lillian Gassie, Naval Postgraduate School, Monterey, CA

Where to use a taxonomy

- Link the taxonomy and indexing
- Always in sync with the industry
- Keep up to date with terminology
- Automatically index the old data
- Filter newsfeeds
- Search using the taxonomy
- File using the taxonomy
- Spell check using the taxonomy
- Link to translation system
- Catalog using the taxonomy

Value of a Taxonomy

- Improves organization & structure
- Facilitates navigation
- Facilitates knowledge discovery
- Reduces effort
- Saves time

Courtesy of Lillian Gassie, Naval Postgraduate School, Monterey, CA

"Taxonomies are better created by professional indexers or librarians than by domain experts."

Taxonomies

- A well formed taxonomy is based on a thesaurus
- Provides a flexible platform for many views of the taxonomy
- Allows fast deployment
- Is the basis for a good
  - knowledge management system
  - search retrieval system
  - portal interface

A quick look behind the scenes

Search thesaurus
Validate term entry
Block invalid terms
Record candidates

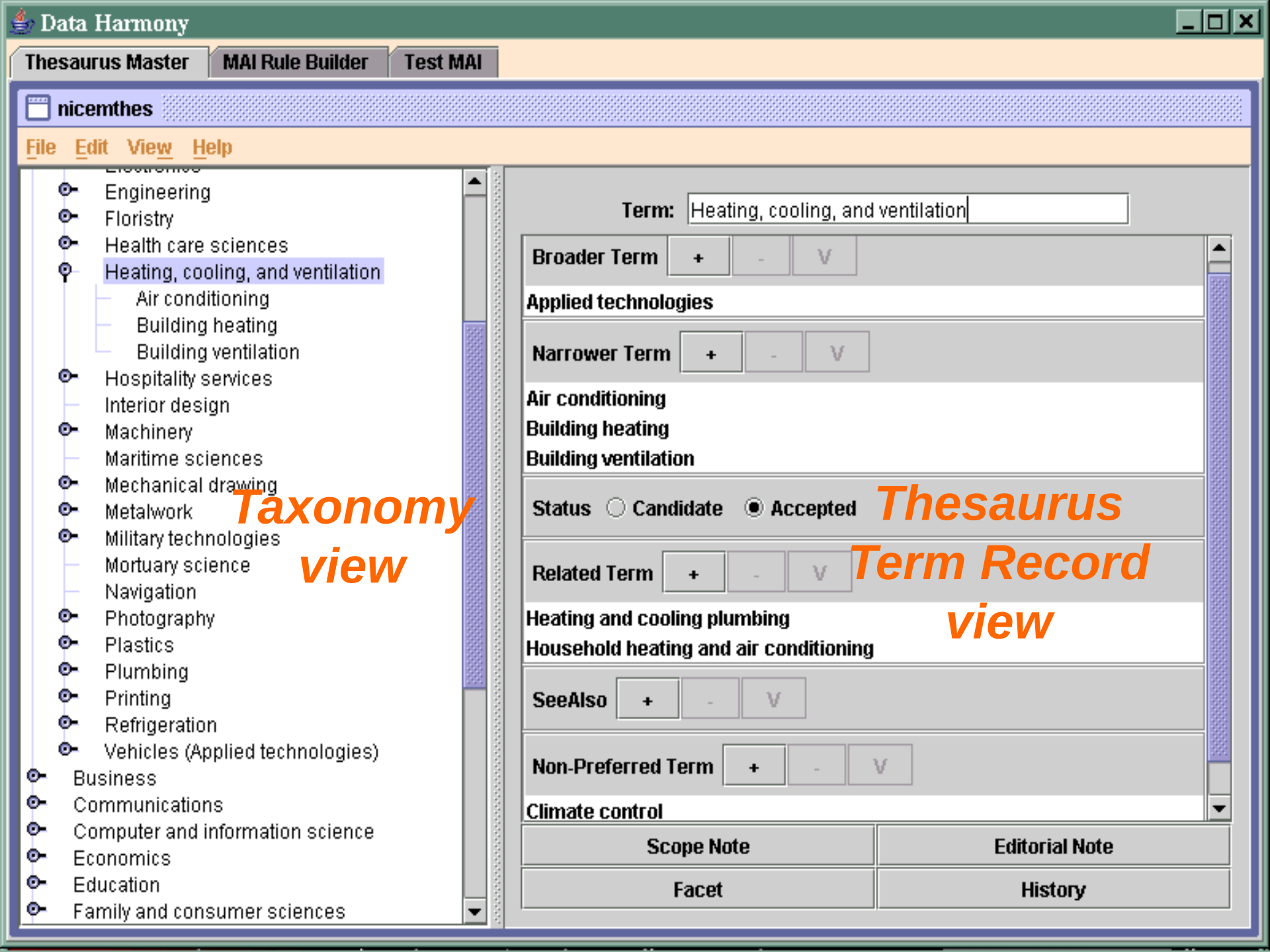
Database Management System

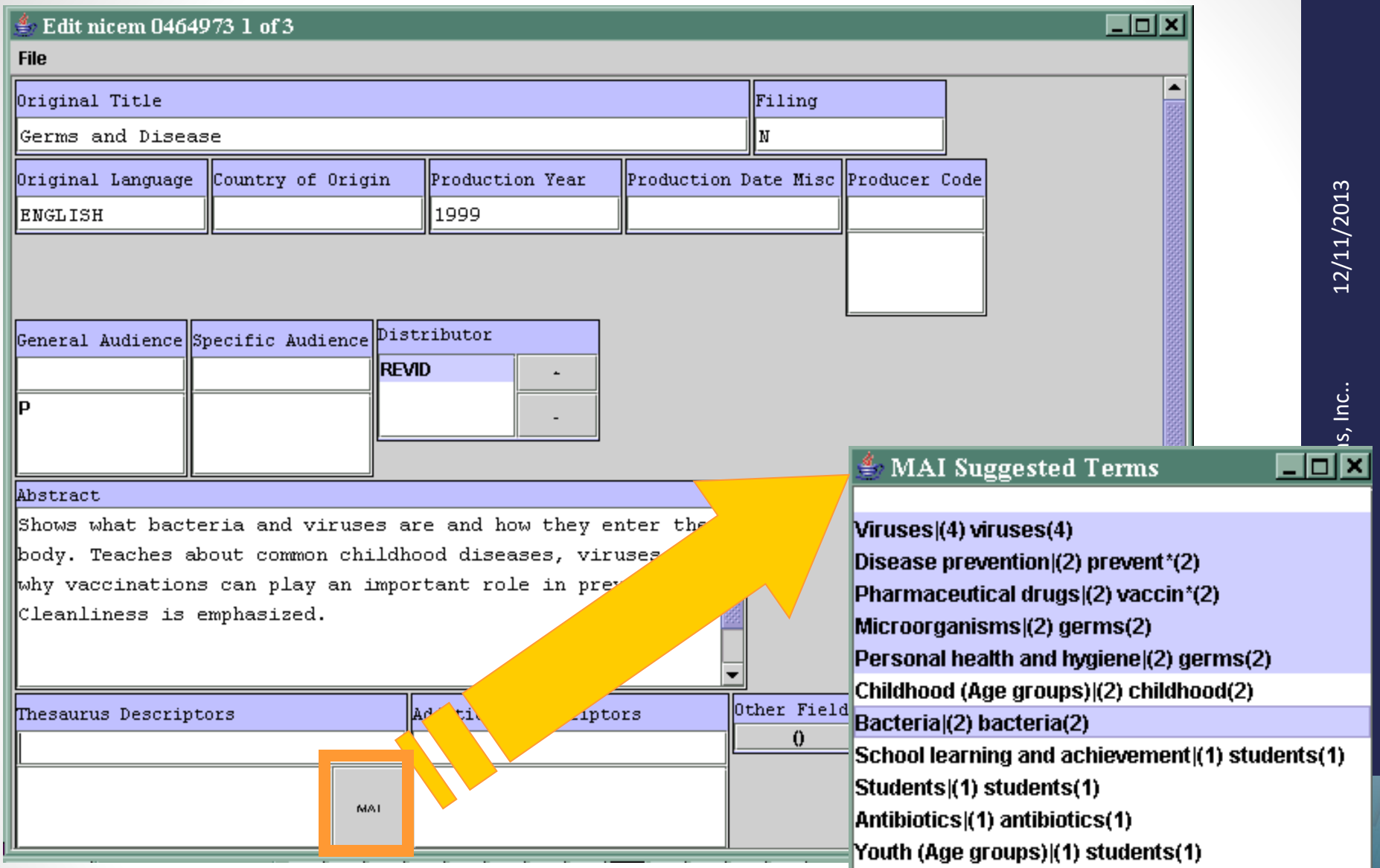
Establish rules for term use
Suggest indexing terms

Thesaurus tool

Validate terms
Add terms and rules
Change terms and rules
Delete terms and rules

Indexing tool


Data Harmony
Thesaurus Master
MAI Rule Builder
Test MAI
nicemthes
File
Edit
View
Help
Engineering
Floristry
Health care sciences
Heating, cooling, and ventilation
Air conditioning
Building heating
Building ventilation
Hospitality services
Interior design
Machinery
Maritime sciences
Mechanical drawing
Metalwork
Military technologies
Mortuary science
Navigation
Photography
Plastics
Plumbing
Printing
Refrigeration
Vehicles (Applied technologies)
Business
Communications
Computer and information science
Economics
Education
Family and consumer sciences
Taxonomy view
Term: Heating, cooling, and ventilation
Broader Term + - V
Applied technologies
Narrower Term + - V
Air conditioning
Building heating
Building ventilation
Status Candidate Accepted
Related Term + - V
Heating and cooling plumbing
Household heating and air conditioning
SeeAlso + - V
Non-Preferred Term + - V
Climate control
Scope Note
Editorial Note
Facet
History
Thesaurus Term Record view



Suggested taxonomy descriptors

Workflow order

- Create the metadata structure
- Gather the locations of the records
- Index in place?
- Point to the data?
- Add metadata to the record?
- Store metadata in separate “table”?
- Use in full text?
- Use in search?
- Link databases and data sets with APIs

APIs and Web Calls

- Link data cross platforms
- Not federated search
- Examples
  - Was: Card catalog or OPACs to books on shelf
  - Now: EBSCO host to Mendeley, Zotero, ResearchGate, Academic.edu, etc.
  - Use an API or web call
- Web calls
  - Call to another web platform
- APIs, written handshakes
  - Z39.50 is one standard for libraries

Discovery (Search)

- Search
  - Free text / full text
  - Fielded / Faceted
- Parts of Search
  - Presentation layer
  - Caching
  - Inverted index

Kinds of search

- Keyword –
  - Autonomy / Verity
- Bayesian –
  - FAST
  - Lucene
- Faceted
  - Endeca
- Boolean
  - Dialog
- Ranking Algorithms
  - Google

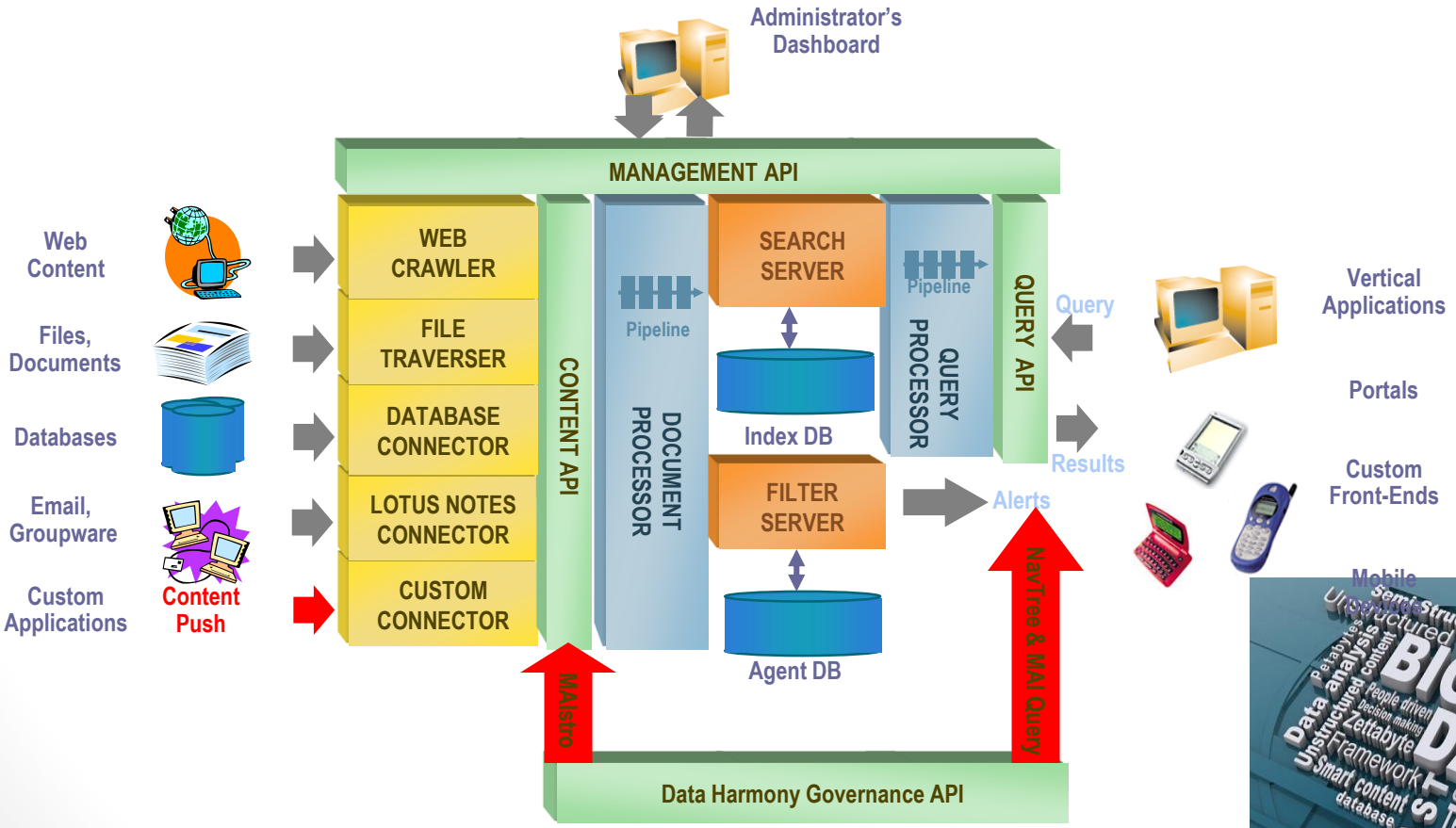
Parts of search

- Query language
- Search technology
- Inverted index
- Ranking algorithms
- Other enhancements

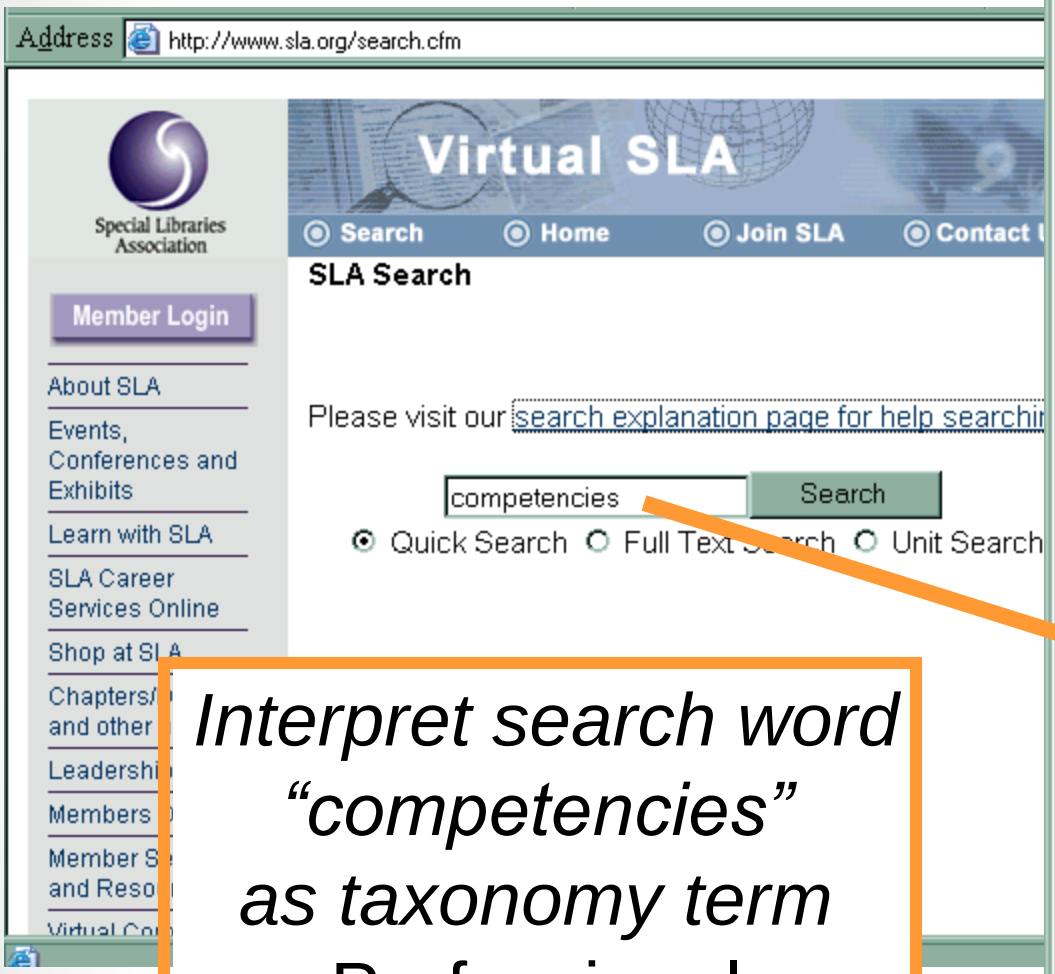
FAST ESP and Data Harmony

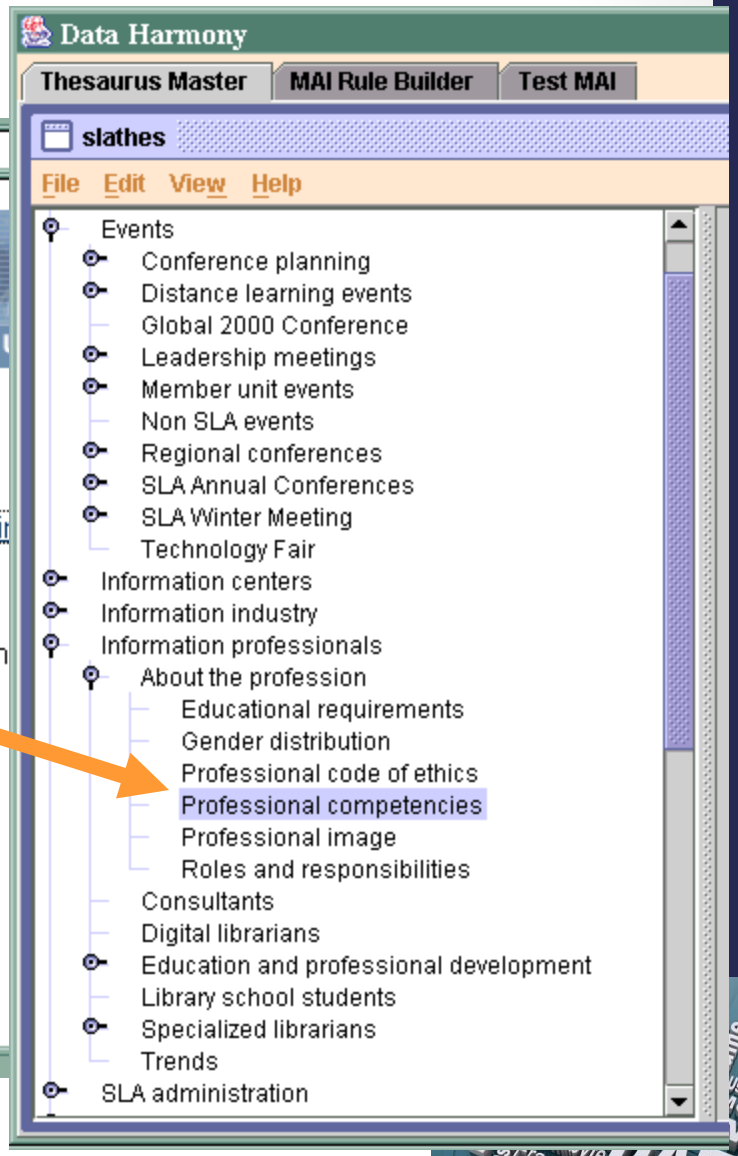
Architecture

Core Architectural Components



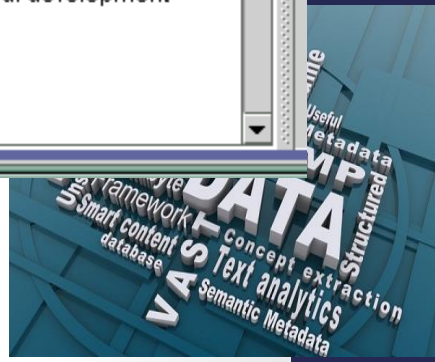
SLA search

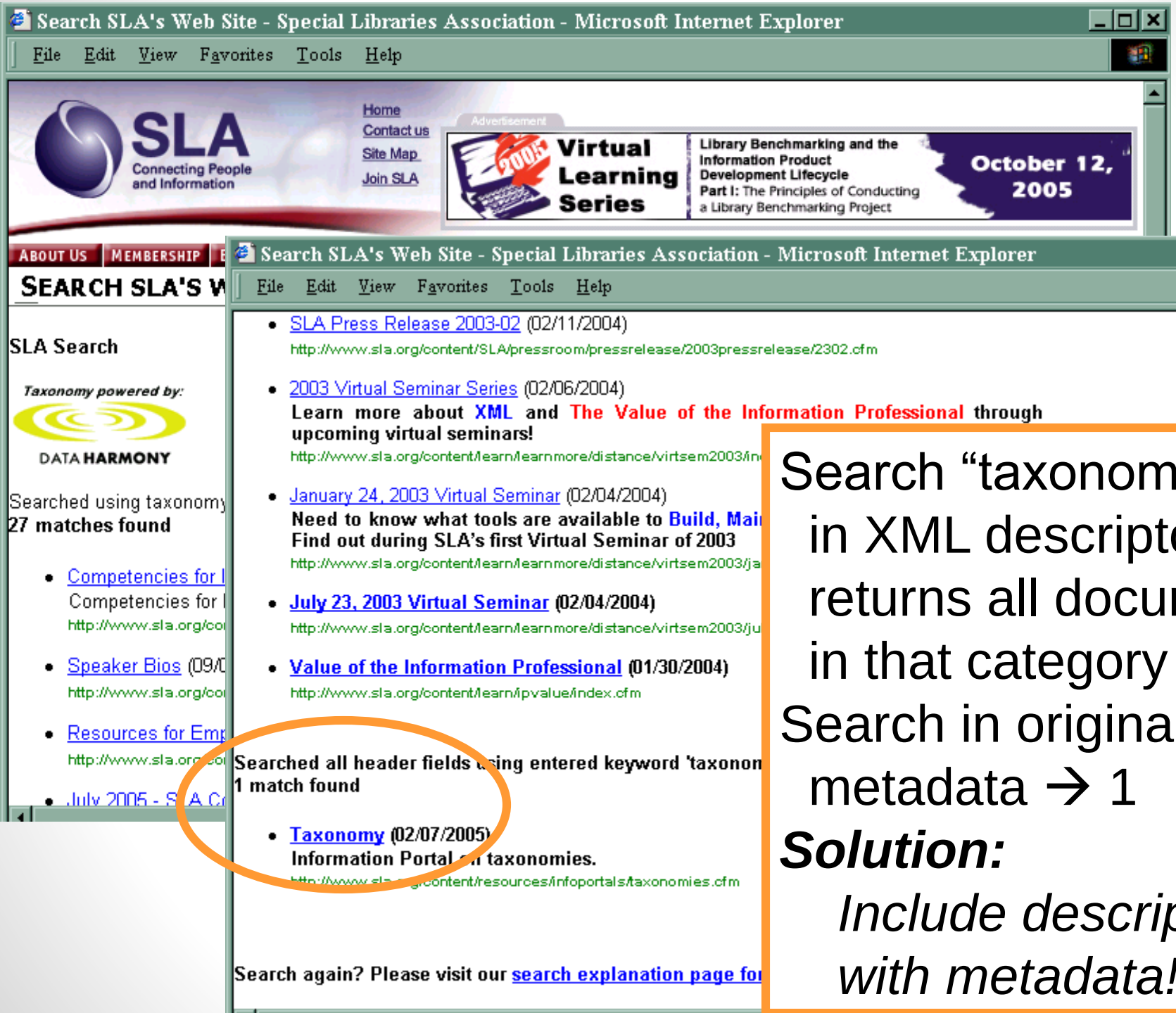




Interpret search word "competencies" as taxonomy term

Professional competencies




Search SLA's Web Site - Special Libraries Association - Microsoft Internet Explorer
File Edit View Favorites Tools Help
SLA
Connecting People and Information
Home
Contact us
Site Map
Join SLA
Advertisement
Virtual Learning Series
Library Benchmarking and the Information Product Development Lifecycle
Part I: The Principles of Conducting a Library Benchmarking Project
October 12, 2005
ABOUT US MEMBERSHIP
SEARCH SLA'S W
SLA Search
Taxonomy powered by:
DATA HARMONY
Searched using taxonomy
27 matches found
Competencies for I
Competencies for I
http://www.sla.org/co
Speaker Bios (09/0
http://www.sla.org/co
Resources for Emp
http://www.sla.org/co
July 2005 - SLA Co
Search SLA's Web Site - Special Libraries Association - Microsoft Internet Explorer
File Edit View Favorites Tools Help
SLA Press Release 2003-02 (02/11/2004)
http://www.sla.org/content/SLA/pressroom/pressrelease/2003pressrelease/2302.cfm
2003 Virtual Seminar Series (02/06/2004)
Learn more about XML and The Value of the Information Professional through upcoming virtual seminars!
http://www.sla.org/content/learn/learnmore/distance/virtsem2003/in
January 24, 2003 Virtual Seminar (02/04/2004)
Need to know what tools are available to Build, Mai
Find out during SLA's first Virtual Seminar of 2003
http://www.sla.org/content/learn/learnmore/distance/virtsem2003/ja
July 23, 2003 Virtual Seminar (02/04/2004)
http://www.sla.org/content/learn/learnmore/distance/virtsem2003/ju
Value of the Information Professional (01/30/2004)
http://www.sla.org/content/learn/ipvalue/index.cfm
Searched all header fields using entered keyword 'taxonon
1 match found
Taxonomy (02/07/2005)
Information Portal on taxonomies.
http://www.sla.org/content/resources/infoportals/taxonomies.cfm
Search again? Please visit our search explanation page for
Search "taxonomy"
in XML descriptor field returns all documents in that category → 27
Search in original metadata → 1
Solution:
Include descriptors with metadata!

Management

- Data management layers
- Curation
- Preservation
- Archiving
- Storage
- Choose tools wisely
- Data mashups

Go – No Go

- Reach 85% precision to launch for productivity - assisted
- Reach 85% for filtering or categorization
  - Sorting for production
- Level of effort to get to 85%
- Integration into the workflow is efficient

Learn to understand the parts

- Information infrastructure
- Information dynamics
- Tensions on data
- Design elements
- How and when to set the data framework
- Identifiers (taxonomies) Glue to hold data together

Services

- Core cross disciplines
- Combo of humans and machines
- What is open
- What needs a gatekeeper
- Store locally
- Store in cloud
- Some things people do better

Benchmarks

- 15 – 20% irrelevant returns / noise
- Amount of work needed to achieve 85% level
- How good is good enough?
  - Satisfice = satisfaction + suffice
  - How good is good enough?
  - How much error can you put up with?
- Hits, misses, noise

ROI Calculations

- Assume – 5000 term thesaurus
  - 1.5 synonyms per terms
  - 7500 terms total
- Assume 85% accuracy
  - Use assisted for indexing
  - Use automatically for filtering
- Assume $55 per hour for staff
- Assume 10000 records for test batch

Co-occurrence - Training sets

- Collect 20 items per thesaurus entry
  - Preferred and non-preferred terms
  - Find records – could be programmatic
  - Need to collect and review about 60 to get 20 appropriate ones
  - Review records – ensure they are correct
  - 3 minutes average per final selected record
  - = 1 hour per term entry
  - 1 minute to review a record (20 resulting records)
  - 7500 terms at one hour per term = 7500 hours
  - Estimated at 7500 hours $55 / hour = $412,500

Rulebuilding - rules

- 80% of rules built automatically
- 7500 x .8 = 6000
- 20% require complex rules
  - Average rule takes 5 minutes
  - (Actually MUCH faster using a rule building tool)
  - 5 x 1500 = 7500 minutes
  - 125 hours x $55 = $6875

ROI - Segments

- Cost of auto system
- Cost of getting system ready
- Ongoing maintenance
- Increased efficiency
- Increased quality of retrieval
- Cost of legacy system maintenance

ROI – Co-occurrence

- Cost of auto system- $500,000 (could be less, but the one used for this study cost this amount)
- Cost of getting system ready
  - Programming support and integration
    - Estimated at 2 weeks programming $100 / hour = $8000
  - Training sets
    - Estimated at 7500 hours $55 / hour = $412,500
    - Possible need to re-run training set several times
- Ongoing maintenance
  - Estimated at 15% of purchase price for license = $75,000
  - Quarterly retraining to keep up with new terms –
  - Training sets for new terms 50 terms per quarter = 200 x $55=$11,000
- Increased efficiency
- Expected accuracy at 60%
- Increased quality of retrieval
- Cost of legacy system maintenance

ROI – Rules system

- Cost of auto system- $60,000
- Cost of getting system ready
  - Programming support and integration
    - Estimated at 2 weeks programming at $100 / hour = $8000
  - Rule building
    - Estimated at 125 hours at $55 / hour = $6850
    - Possible need to re run training set several times
- Ongoing maintenance
  - Estimated at 15% of purchase price for license = $9000
  - Rule building for new terms, 50 terms per quarter
    - 200 terms x .8 = 160 automatic
    - 40 at 5 minutes per term = 200 minutes /60 = 3.33 hours x $55 = $183
- Increased efficiency
- Expected accuracy at 60%
- Increased quality of retrieval
- Cost of legacy system maintenance

ROI - Rules

- Year one
  - $60,000 + $8,000 + $6,850 = $74,850
- Years thereafter
  - $9000 + $183 = $9183
- 85% accuracy

ROI – Co-occurrence

- Year one
  - $500,000 + $8,000 + $412,500 = $920,500
- Years thereafter
  - $75000 + $11,000 = $86,000
- 60% accuracy

Summary on ROI

- Novelty detection requires inference techniques
- Is needed to discover new terms
- Controlled vocabulary / thesaurus can be more controlled and accurate approach
- Combination of the two would be best
  - Controlled – use rules
  - New terms - use inference

Fully integrated with MOSS
The Workflow

Gather source data → Tag and Create metadata → Put in data base with tags → Build Search inverted index → Create user interface

Client Data
Full Text
HTML, PDF,
Data Feeds, etc.

Automatic Summarization
Machine Aided Indexer (M.A.I.™)
Inline Tagging
Metadata and Entity Extractor
Thesaurus Master

Client taxonomy

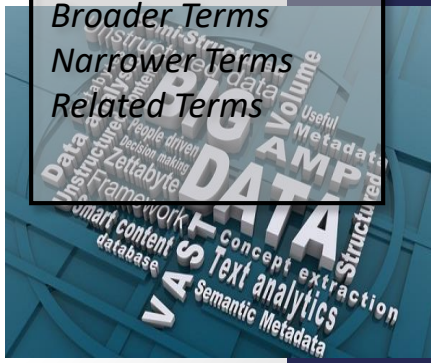
Database Repository

Search Software

Search Presentation Layer

Increases accuracy

Browse by Subject
Auto-completion
Broader Terms
Narrower Terms
Related Terms



Inline Tagging

oil rises as U.S. stimulus hopes outweigh weak demand **oil prices** rose on Thursday as hopes that the White House would move quickly on an economic stimulus package outweighed flagging demand and rising **inventories** in the world's top **consumers**. U.S. **crude** settled 12 cents higher at $43.67 a barrel, after falling as low as $40.41 earlier. London Brent settled at $45.39 a barrel, up 37 cents. Earlier in the day, **crude prices** had dropped after a U.S. government report showed that **crude oil**, **gasoline** and distillate fuels rose last week as demand for fuels weakened again. The U.S. stock **market** pared early losses after a spokesman said President Obama's administration is committed to moving as quickly as possible on an economic stimulus plan. "**crude** is still resilient, despite the big build you have seen in the EIA data. I have a feeling that there will soon be a rebound in the stock **market** and that will spill over to the **energy market**s," said Mark Waggoner, president of Excel **Futures** in Huntington Beach, **California**. "In the last week, I've seen **invest**or confidence improving, with the new Obama administration now installed ... I feel that if this confidence continues, it will spill into more (consumer) buying and that will improve demand for **gasoline** and other **energy product**s," he added. **oil** fell in early **trade** after U.S. **energy** Information Administration data showed that **crude oil inventories** jumped 6.1 million barrels last week, well above expectations for a build of 1.4 million barrels.

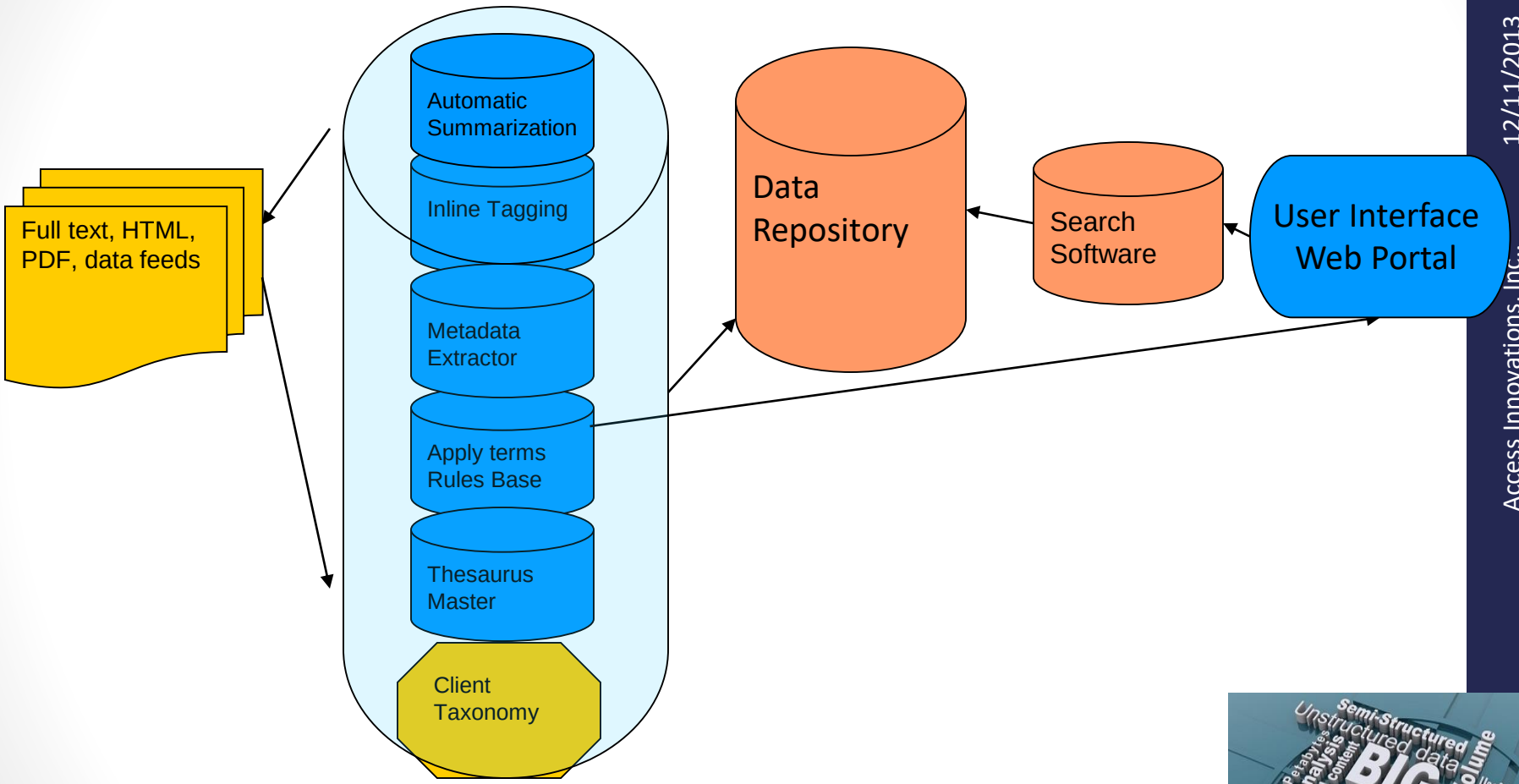
MAIstroTerm : Crude oil (Commodity markets)
MAIstroTerm : Petroleum resources
MAIstroTerm : Oil (Fuels)

MAIstroTermList

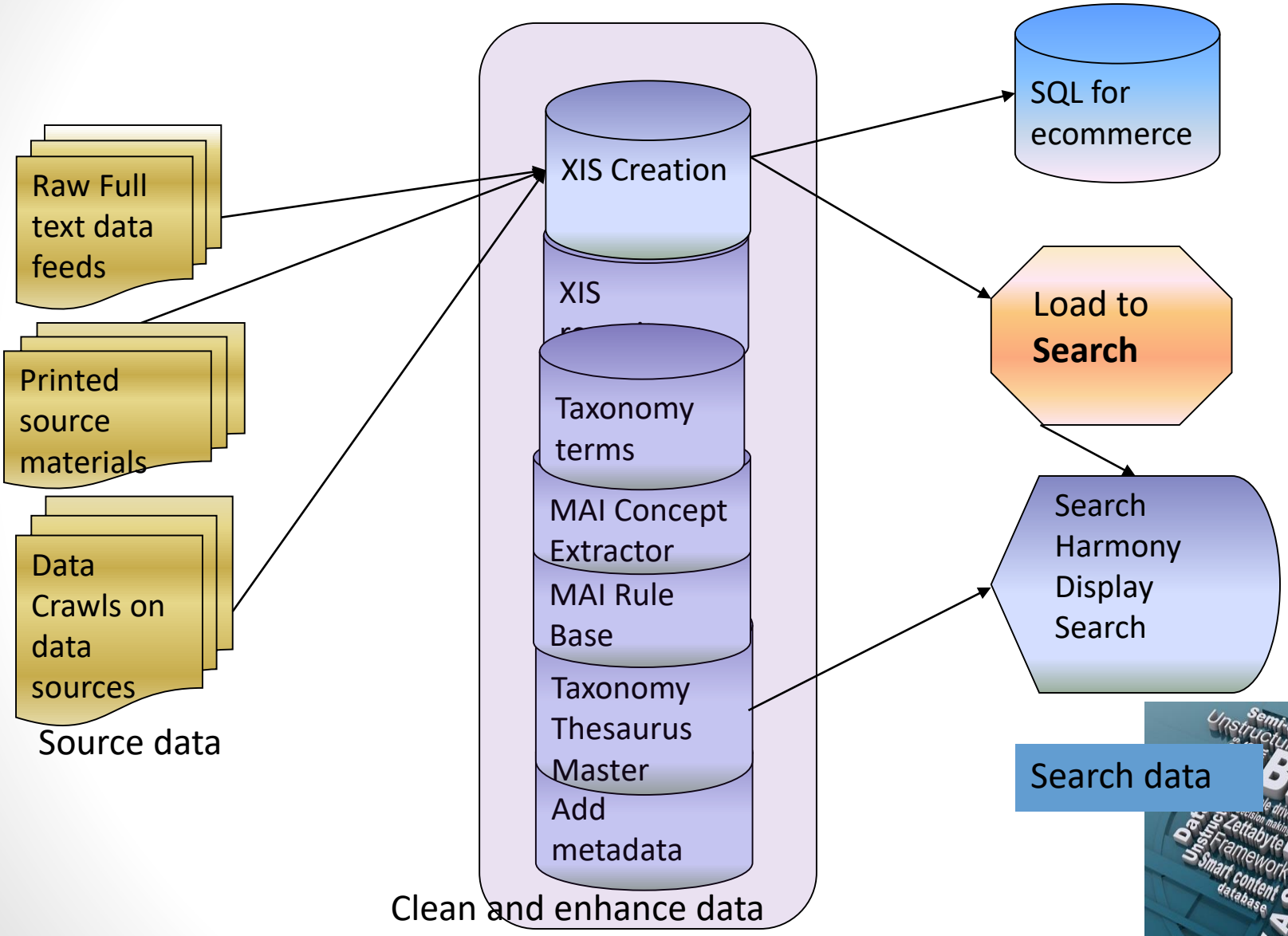
Term Name="**Crude oil (Commodity markets)**" NumberOfTimes="3"
Term Name="**Petroleum resources**" NumberOfTimes="**3**"
Term Name="**Oil (Fuels)**" NumberOfTimes="**3**"
Term Name="**Energy resources (Commodity markets)**" NumberOfTimes="**3**"
Term Name="**Product inventories**" NumberOfTimes="**3**"
Term Name="**Gasoline**" NumberOfTimes="**2**"
Term Name="**Prices**" NumberOfTimes="**2**"
Term Name="**Commodity market prices**" NumberOfTimes="**2**"
Term Name="**Stock market**" NumberOfTimes="**2**"
Term Name="**California**" NumberOfTimes="**1**"
Term Name="**Futures (Investments)**" NumberOfTimes="**1**"
Term Name="**Commerce**" NumberOfTimes="**1**"
Term Name="**Money and banking**" NumberOfTimes="**1**"

- Shows the exact point where the concept is mentioned
- Mouse-over to view the term record
- Statistical summary, showing the number of times each term is mentioned in the article

Semantic Process



Database Plus Search Workflow



Sample DOCUMENT

Creating an Inverted File Index

Outline of Presentation

1. Define key terminology
2. Thesaurus tools
   - Features
   - Functions
3. Costs
   - Thesaurus construction
   - Thesaurus tools
4. Why & when?

The terms from the “outline”

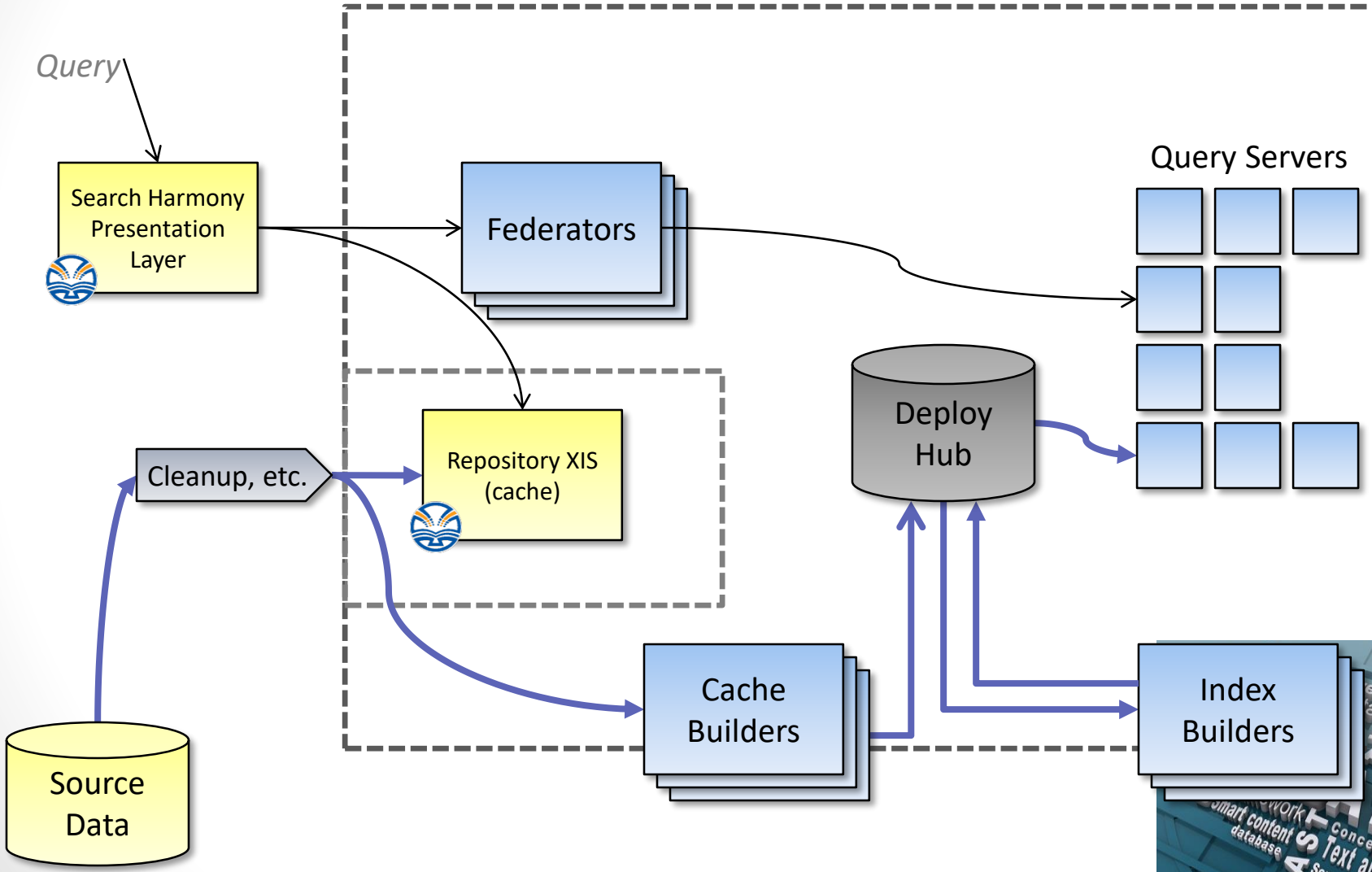
&
1
2
3
4
construction
costs
define
features
functions

key
of
outline
presentation
terminology
thesaurus
tools
when
why

Complex inverted file index Placement location

& - Stop
1 - Stop
2 - Stop
3 - Stop
4 - Stop
construction - L7, P2, SH
costs - L6, P1, H
define - L2, P1, H
features - L4, P1, SH
functions - L5, P1, SH
key - L2, P2, H
of - Stop
outline - L1, P1, T
presentation - L1, P3, T
terminology - L2, P3, H
thesaurus - (1) - L3, P1, H
(2) - L7, P1, SH
(3) - L8, P1, SH
tools - (1) - L3, P2, H
(2) - L8, P2, SH
when - L9, P3, H
why - L9, P1, H

Complex Search Farm



What happens at the search presentation layer?

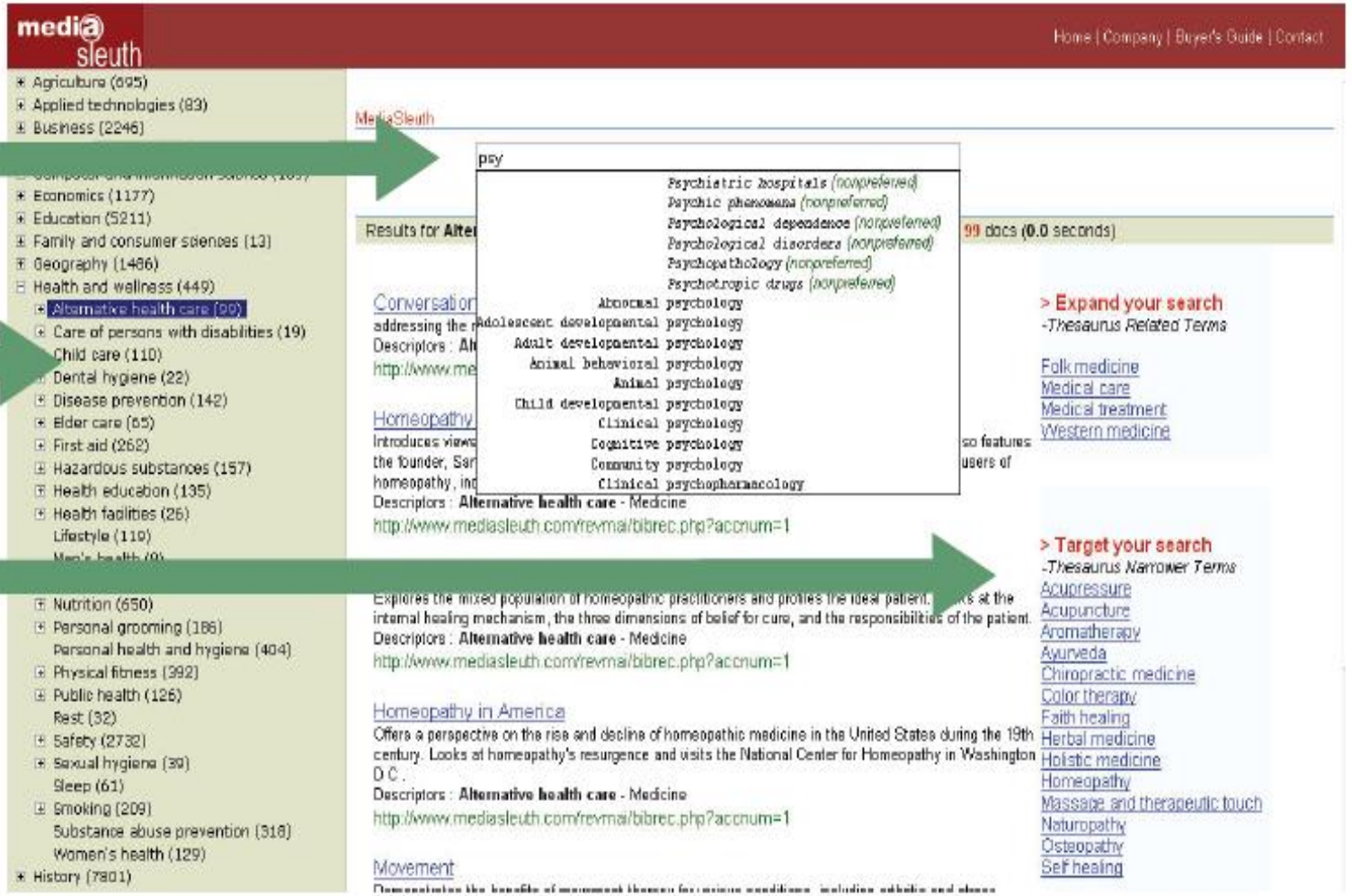
- That is what librarians usually look at.
- What are the options coming?
- How can we encourage useful changes?

Semantic search options

Autocompletion using controlled terms

Full taxonomy view for Navigation and Browsing

Related, Broader, and Narrower Terms



OPEN ACCESS PEER-REVIEWED
RESEARCH ARTICLE

Predation by Bears Drives Senescence in Natural Populations of Salmon

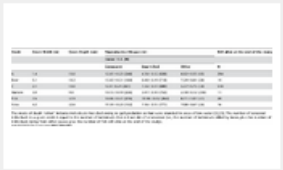
Stephanie M. Carlson, Ray Hilborn, Andrew P. Hendry, Thomas P. Quinn

Published: Dec 12, 2007 • DOI: 10.1371/journal.pone.0001286

Abstract

Classic evolutionary theory predicts that populations experiencing higher rates of environmentally caused ("extrinsic") mortality should senesce more rapidly, but this theory usually neglects plausible relationships between an individual's senescent condition and its susceptibility to extrinsic mortality. We tested for the evolutionary importance of this condition dependence by comparing senescence rates among natural populations of sockeye salmon (*Oncorhynchus nerka*) subject to varying degrees of predation by brown bears (*Ursus arctos*). We related senescence rates in six populations to (1) the overall rate of extrinsic mortality, and (2) the degree of condition dependence in this mortality. Senescence rates were determined by modeling the mortality of individually-tagged breeding salmon at each site. The overall rate of extrinsic mortality was estimated as the long-term average of the annual percentage of salmon killed by bears. The degree of condition dependence was estimated as the extent to which bears killed salmon that exhibited varying degrees of senescence. We found that the degree of condition dependence in extrinsic mortality was very important in driving senescence: populations where bears selectively killed fish showing advanced senescence were those that senesced least rapidly. The overall rate of extrinsic mortality also contributed to among-population variation in senescence-but to a lesser extent. Condition-dependent susceptibility to extrinsic mortality should be incorporated more often into theoretical models and should be explicitly tested in natural populations.

Figures


Subject Areas

Bears
Senescence
Predation
Fishes
Freshwater fish
Death rates
Marine fish
Animal sexual behav...

OPEN ACCESS PEER-REVIEWED
RESEARCH ARTICLE

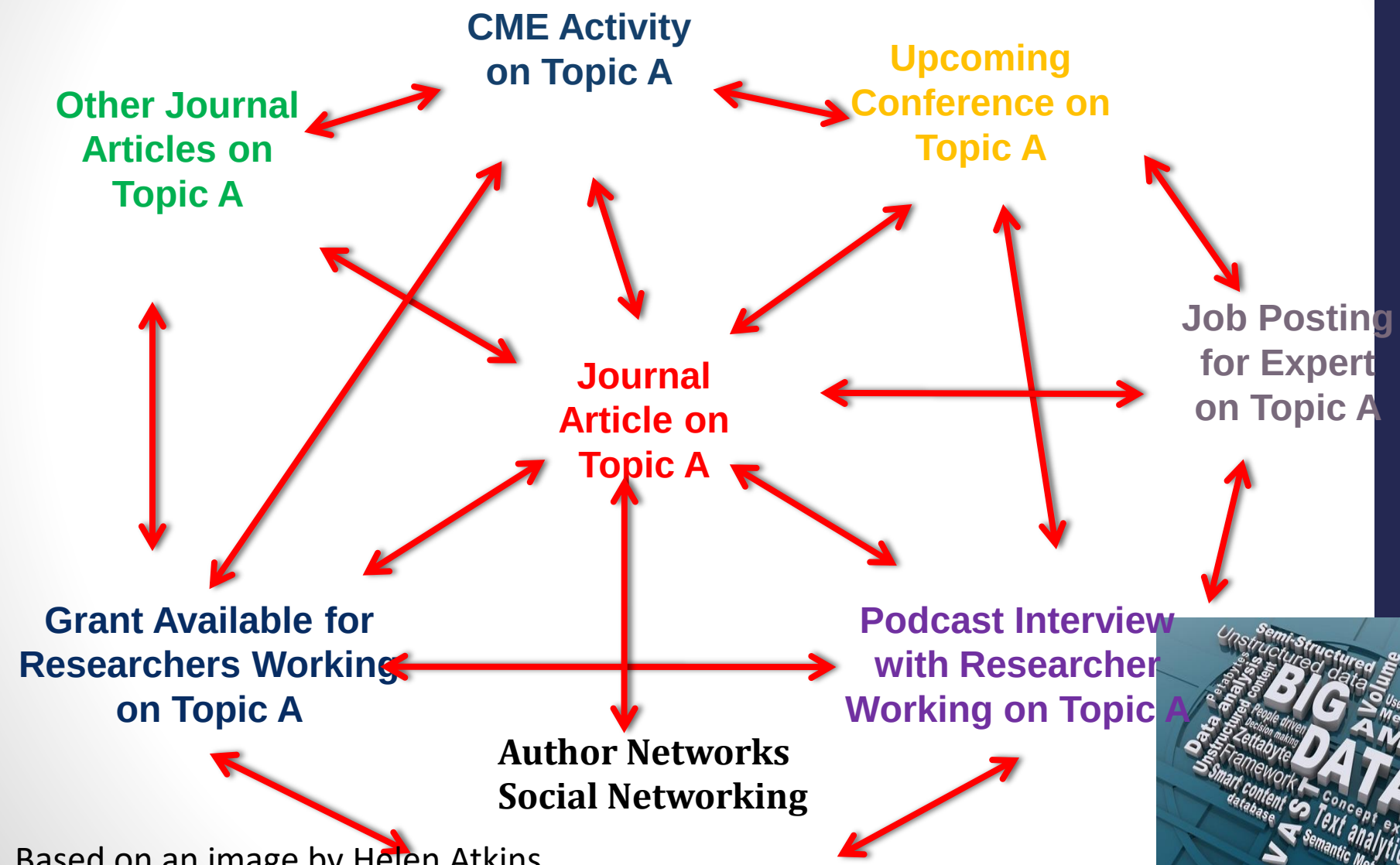
Predation by Bears Drives Senescence in Natural Populations of Salmon

Stephanie M. Carlson, Ray Hilborn, Andrew P. Hendry, Thomas P. Quinn

Published: Dec 12, 2007 • DOI: 10.1371/journal.pone.0001286

Abstract

Data linked and served by metadata



Based on an image by Helen Atkins

American Association for Cancer Research

Cancer Epidemiology Biomarkers & Prevention
Vol. 12, 161-164,
February 2003

Short Communications

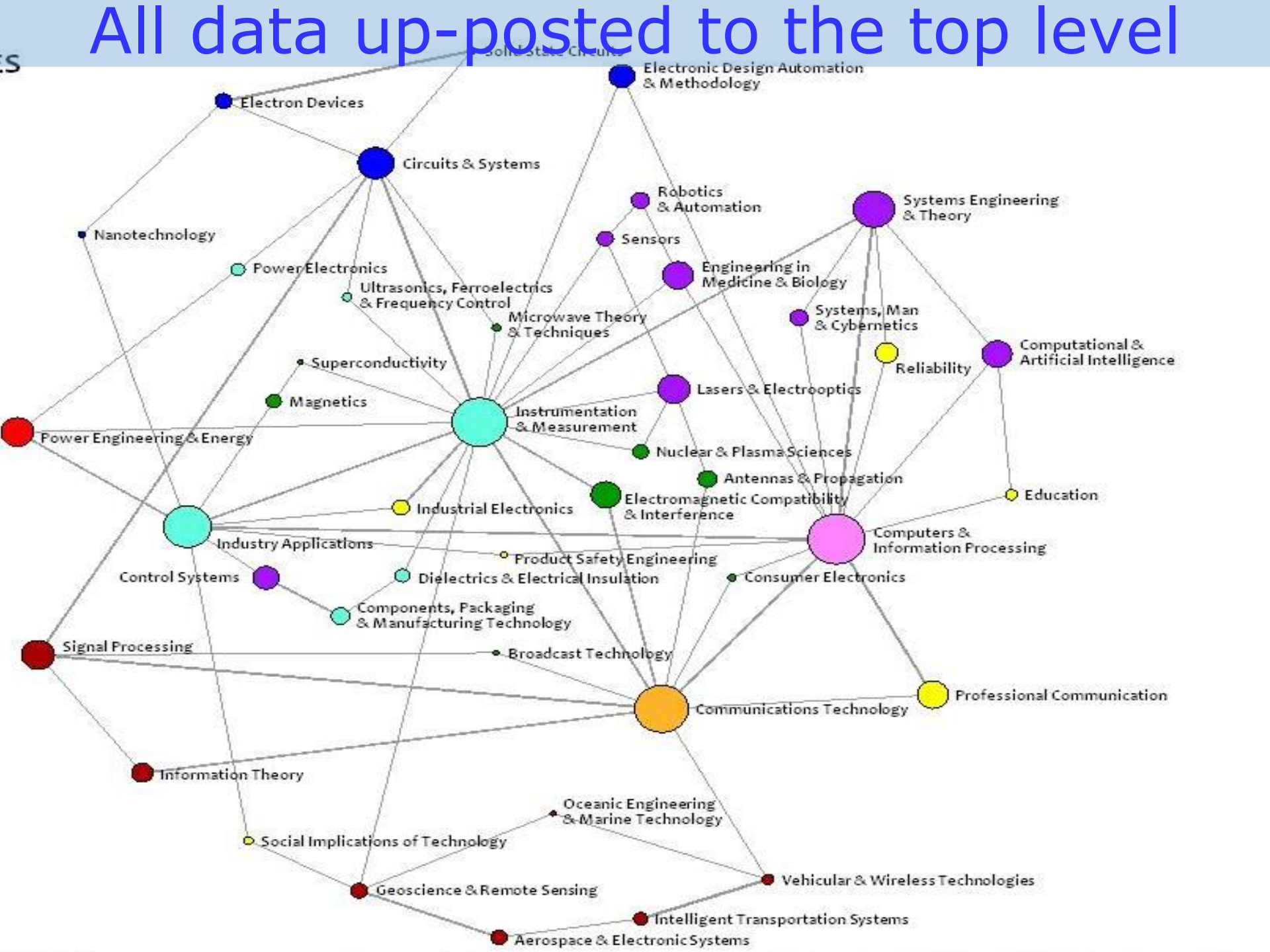
Alcohol, Folate, Methionine, and Risk of Incident Breast Cancer in the American Cancer Society Cancer Prevention Study II Nutrition Cohort

Heather Spencer Feigelson[1], Carolyn R. Jonas, Andreas S. Robertson, Marjorie L. McCullough, Michael J. Thun and Eugenia E. Calle
Department of Epidemiology and Surveillance Research, American Cancer Society, National Home Office, Atlanta, Georgia 30329-4251

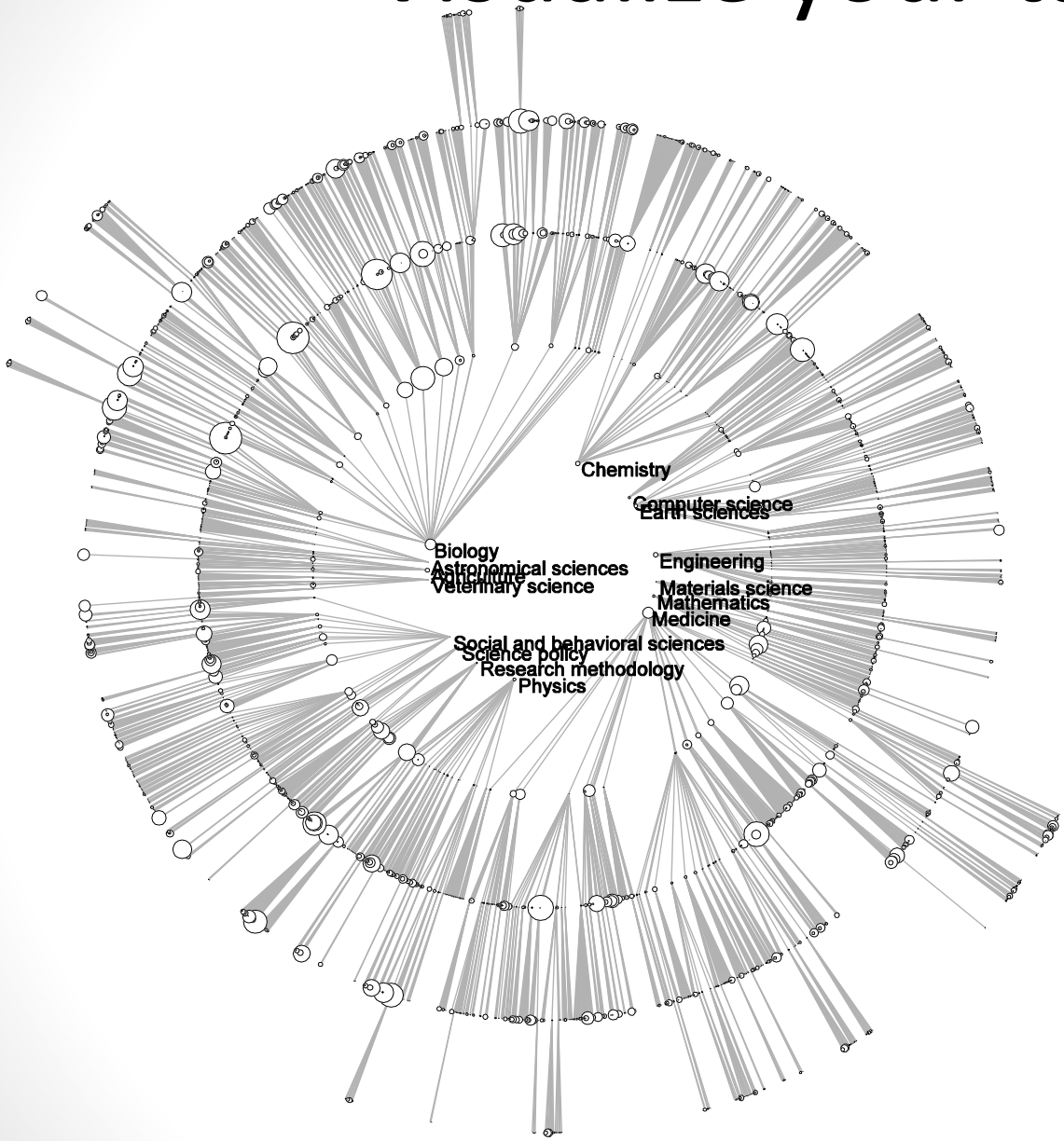
Recent studies suggest that the increased risk of breast cancer associated with alcohol consumption may be reduced by adequate folate intake. We examined this question among 66,561 postmenopausal women in the American Cancer Society Cancer Prevention Study II Nutrition Cohort.

After Helen Atkins

All data up-posted to the top level

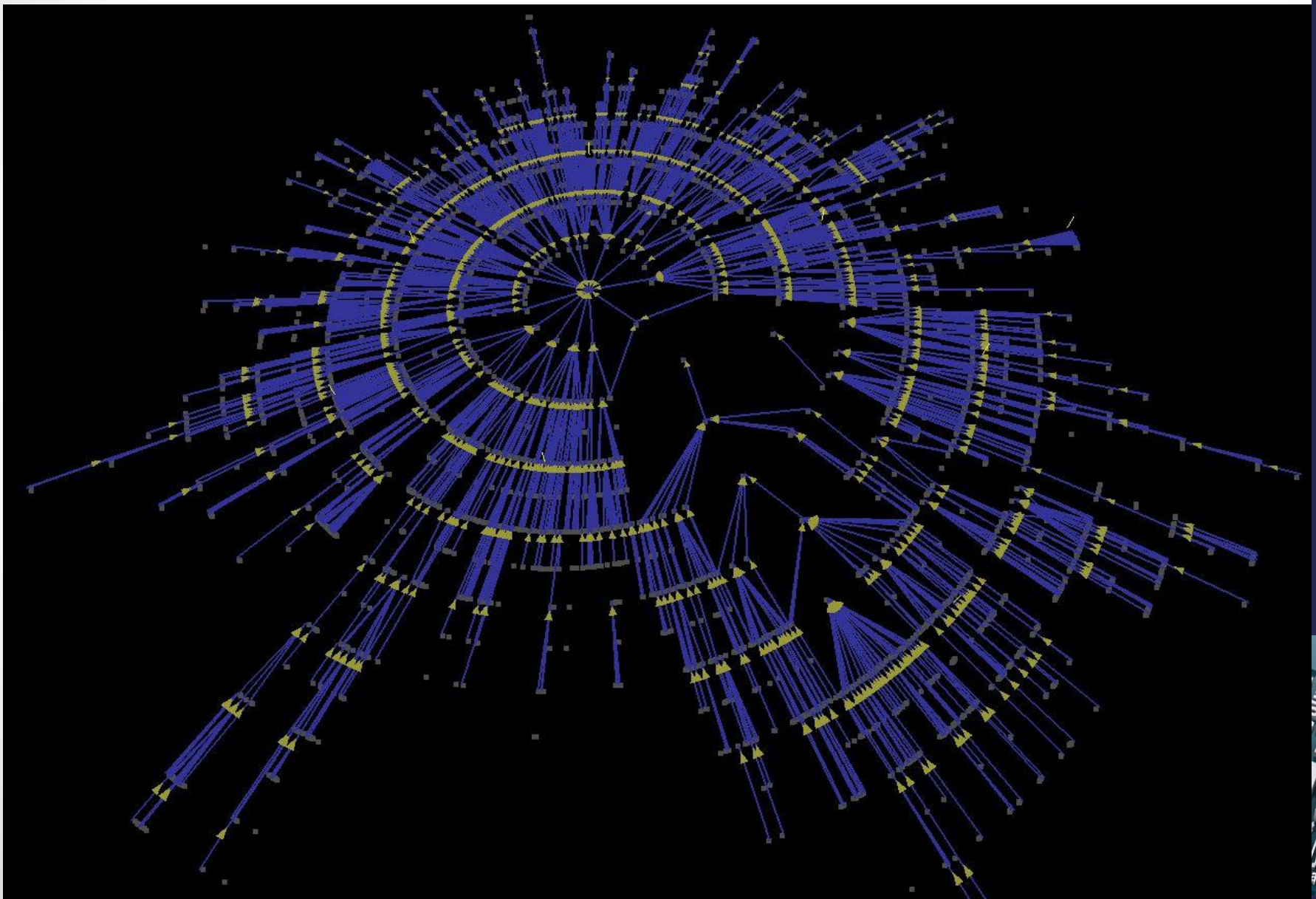


Visualize your tagged data

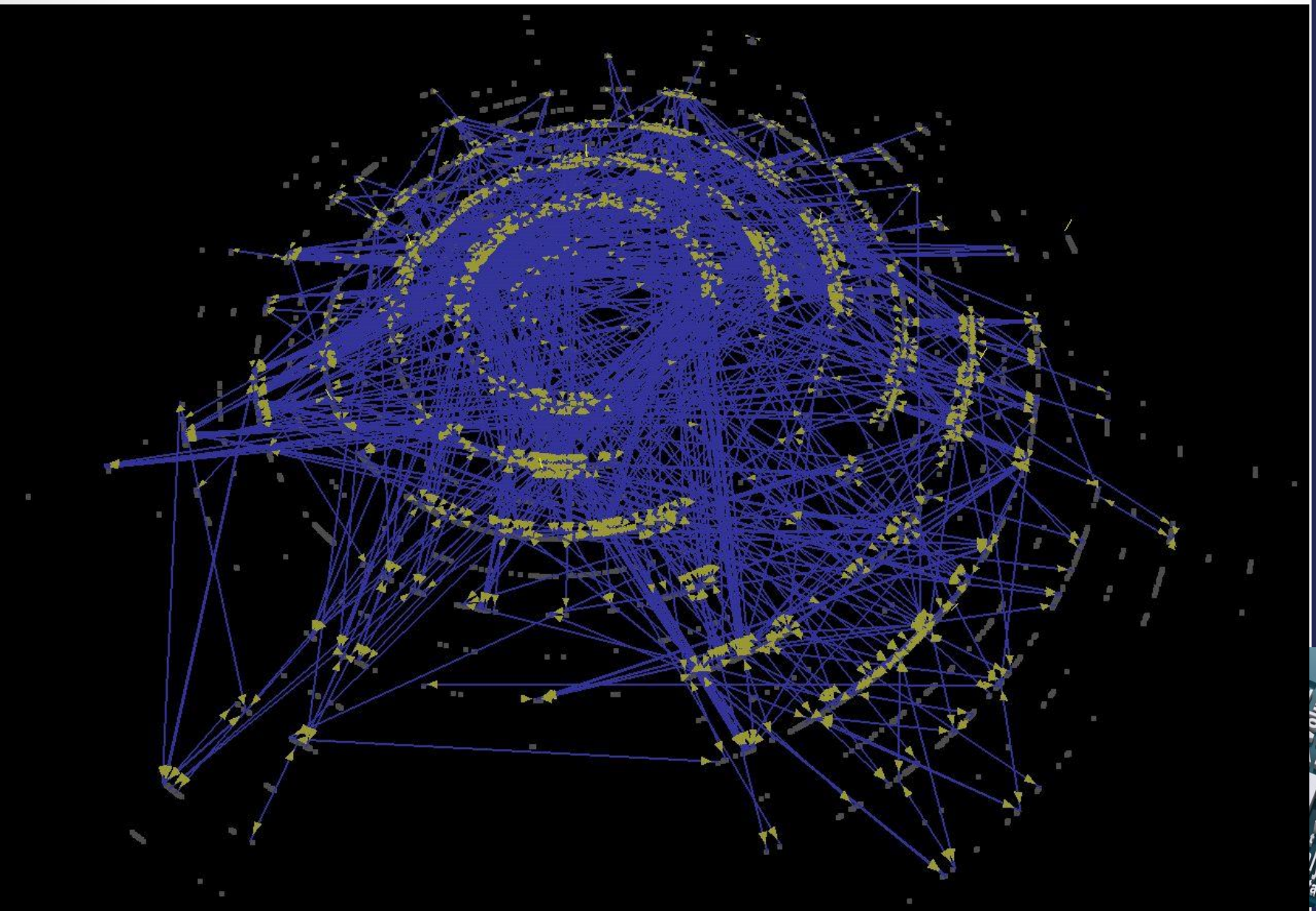


This is a radial graph of "plosthes". The number of records for which each index term occurs is reflected by circle sizes.

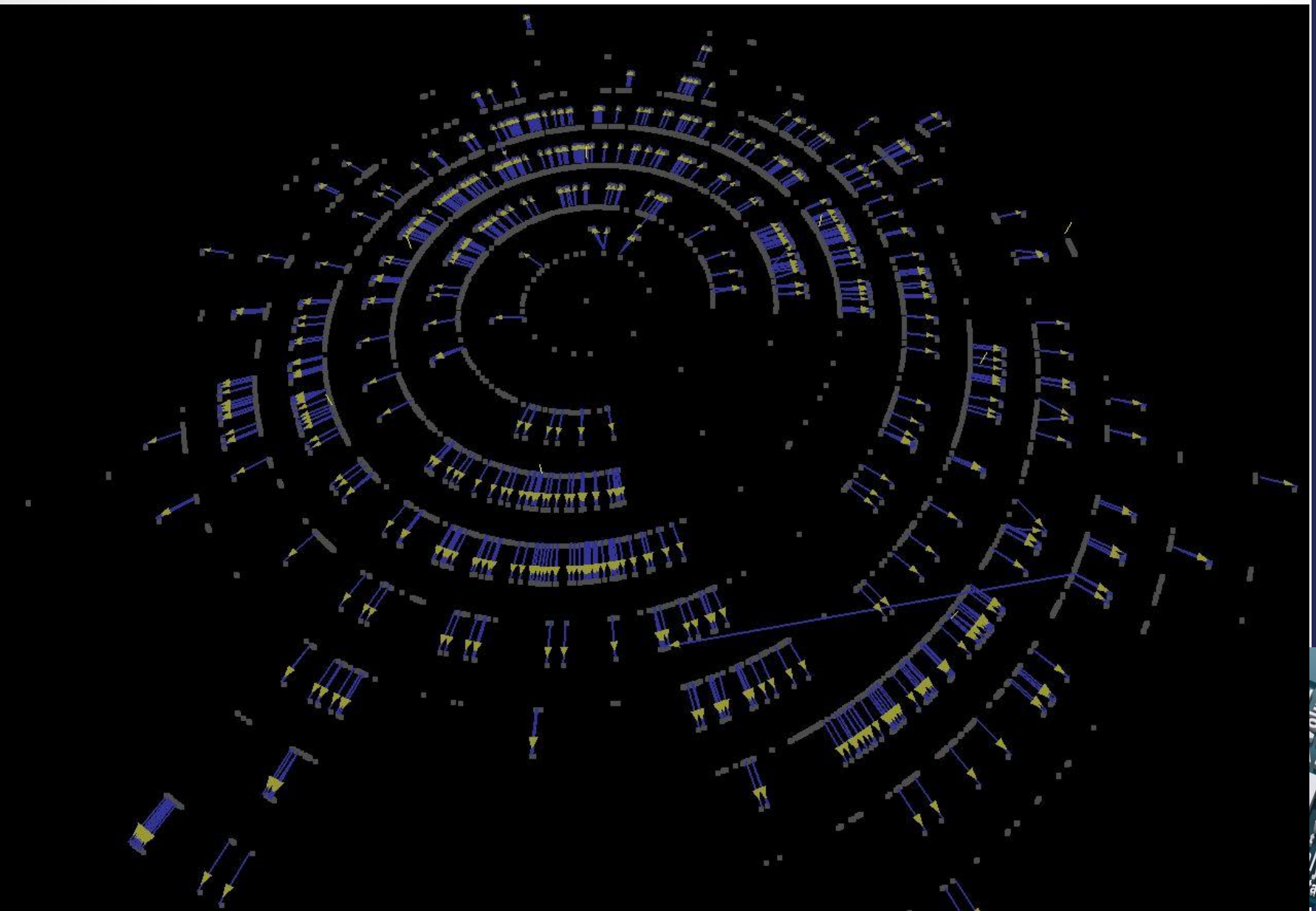
Semantic Hierarchy → browsable tree



Related Terms → semantic web



Synonyms → search foundation



Load to a visualization program such as Prefuse

prefuse
INFORMATION VISUALIZATION TOOLKIT

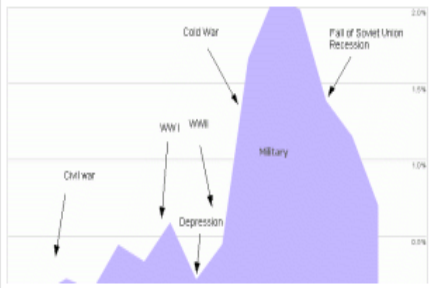
Home | Download | Gallery | Documentation | FAQ

Search

visualization gallery

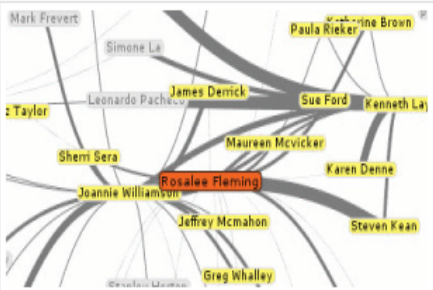
DocuBurst by Christopher Collins

Uncertainty Lattices by Christopher Collins

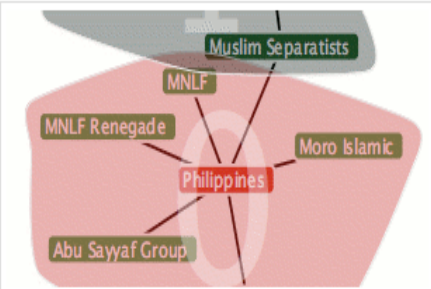


sense.us by Heer, Viégas, and Wattenberg

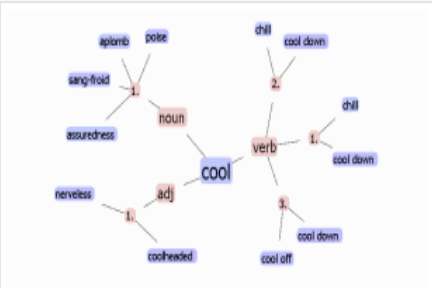

StudiAnalyse by Christoph Gerstle and Florian Moritz



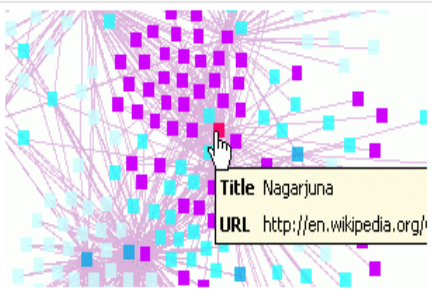
Enron Explorer by Trampoline Systems



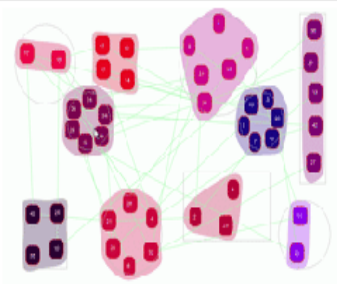
Social Action by Adam Perer

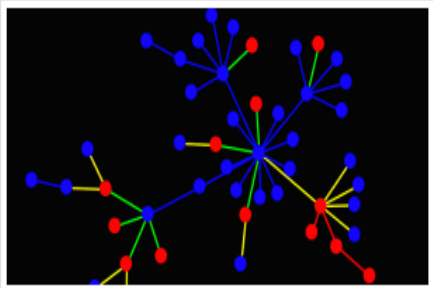


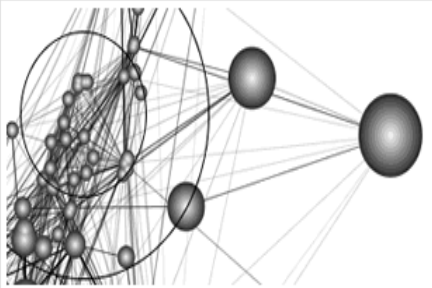
Nearword by Gregory Vaughan

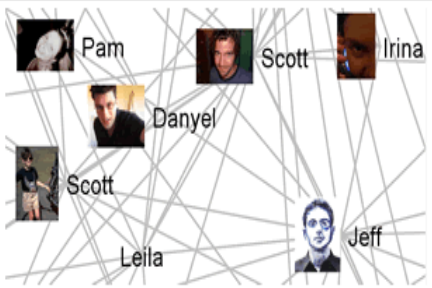


34all by Martin Dudek


Zone Manager by Martin Dudek


timeVis by Can Altineller


Small-World Networks by Stephen Frowe Ingram



Vizster by Jeffrey Heer and danah boyd

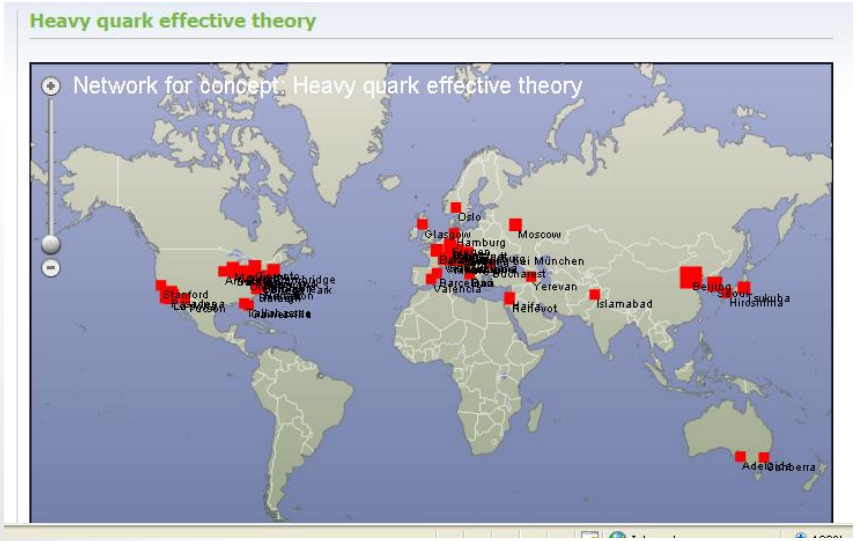
Data Mashups

- Drawing together information from several sources
- Overlay on additional surfaces – like maps
- Fun distribution of data
- Cornell School of Ornithology
  - Migration patterns
  - YouTube
  - Citizen Science

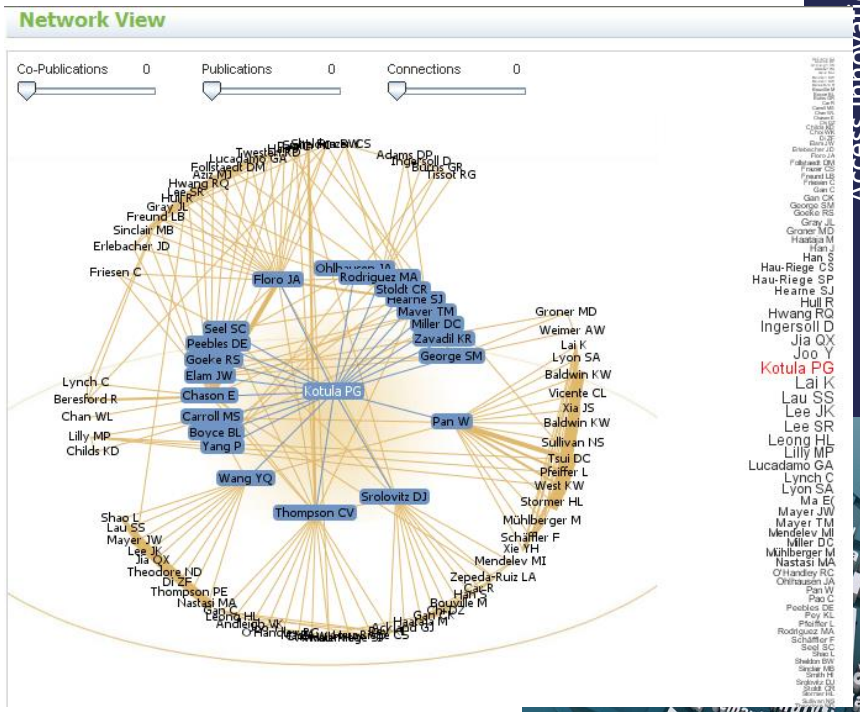
Scientific social networking based on metadata

- Idea has been here - Who is citing who like
  - ISI does it with references
  - AIP's UniPHY does it using semantics
- Expand your options using
  - good metadata and descriptors

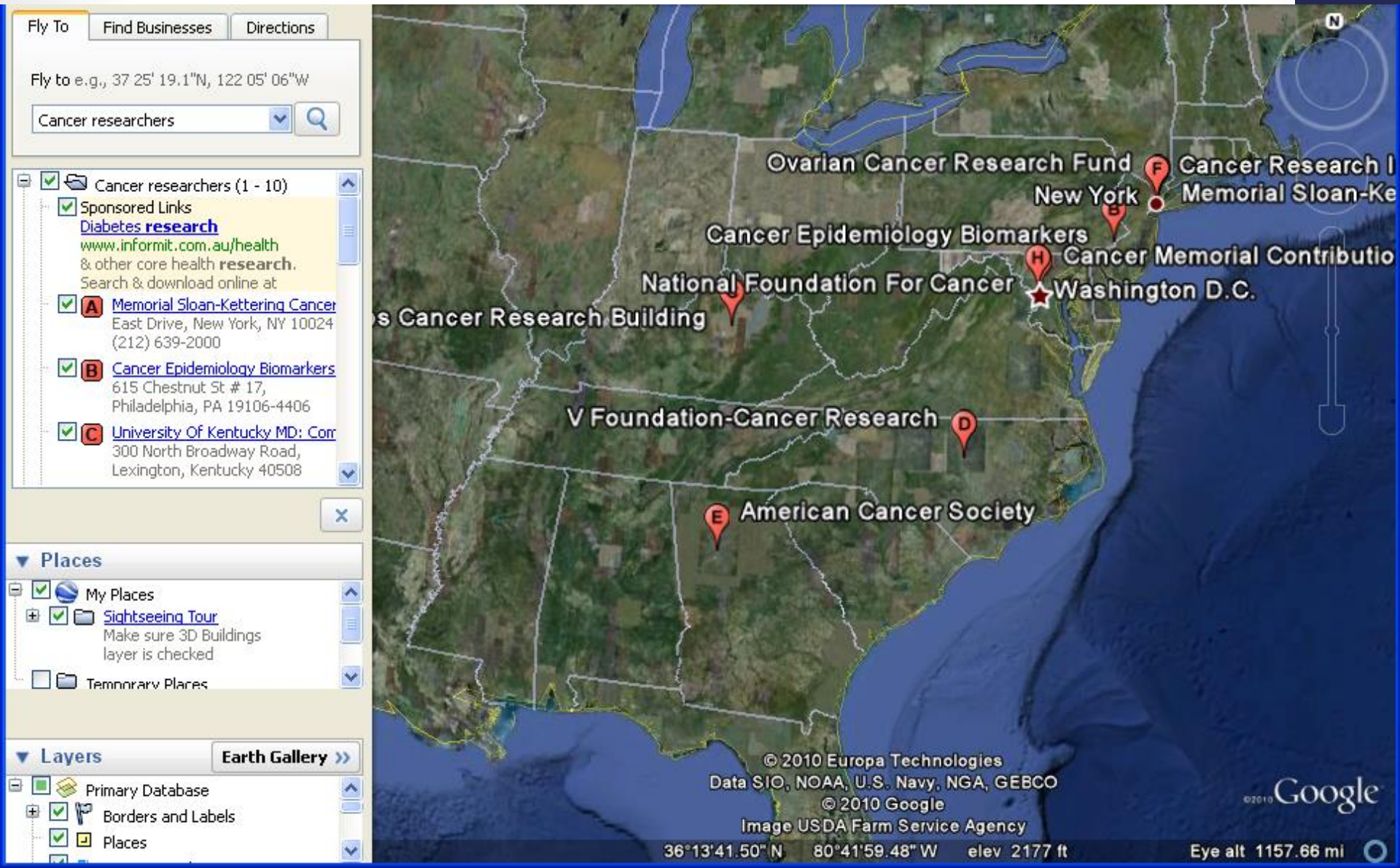
Map who is working in the field and where



See the authors connections



Authors at a Place



Future?

- Document systems replaced
- New infrastructure
- Institutional repositories not good long term for Big Data
- Need to operate at scale
- Integrated, ecosystem to infrastructure
- Replace customized human mediated
  - With interpretive layer – computer assisted
- Links to data
- Open data / supplemental data
- Too much is not enough!

Differentiation

- Not by size of collection
- By the services they offer
- More services more competitive
- Spreadsheet science
  - Go down hall and ask
- Big Data
  - Too many people to ask
  - Lose provenance
  - Hard to differentiate the layers

New ways of doing things

- People, data usage
- Preservation and discovery
- Publishers' content, not format
- Visualization of Big Data sets
- Reproducibility of research
- Shared knowledge
- Open peer review
  - Researcher.org
  - Galaxy zoo
  - Zooniverse
  - 900,000 citizen scientists
  - Wikipedia

Few have witnessed what you're about to see

Experience a privileged glimpse of the distant universe as observed by the SDSS, the Hubble Space Telescope, and UKIRT

GALAXY ZOO QUENCH

We are trying something new! Come help us understand a very specific type of galaxy and experience science from start to end. Take part

Classify Galaxies

To understand how galaxies formed we need your help to classify them according to their shapes. If you're quick, you may even be the first person to see the galaxies you're asked to classify.

Begin Classifying

Skills we bring

- Librarians
  - Weeding
  - Redefining collections
  - Collection development
- Skills to apply to Big Data
  - Vocabulary development
  - Search expertise
  - Reference ability
- What to throw away
- What to keep for discoverability
- Need metadata to preserve
  - Better discoverability
  - Easier preservation
  - Keep the data provenance

We covered

- Big Data – What is it?
- New Government Initiative
- Content organization
- Discovery (Search)
- Management
- Skills we bring
- Examples of what we can do

Who are we?

- Access Innovations –
  - Semantic Enrichment Services Provider
  - We change search to find!
- Creator of Data Harmony tools
- More than 600 taxonomies created
- More than 2000 engagements
- Financed by Sweat, Persistence, and Good Cash Flow Management
- Accurate, on time, under budget!

Slides at www.accessinn.com/presentations

Thank you

- Marjorie Hlava
  - mhlava@accessinn.com
- +1-505-998-0800
- Access Innovations/Data Harmony/Access Integrity
- Taxodiary.com (blog)
- Taxobank.org (reference tool for taxonomists)