

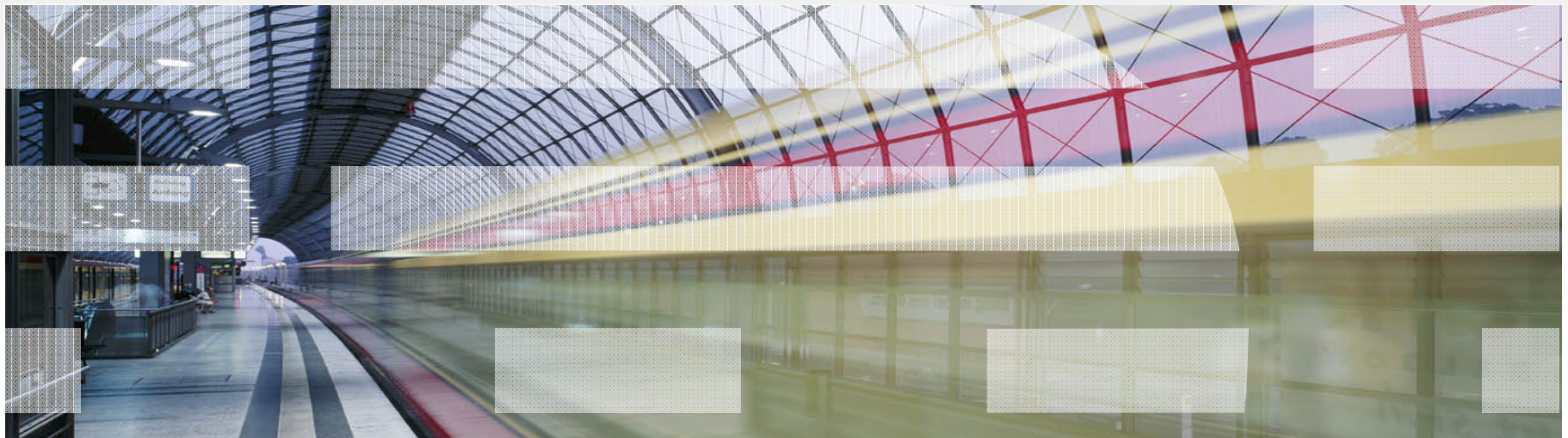


ELSEVIER

Management of Taxonomies for Search, CMS, and Semantic Processing

Presented to San Francisco DAMA, Feb. 9, 2011

Dr. Ron Daniel, Jr. Elsevier Labs



ELSEVIER

Building Insights. Breaking Boundaries.™

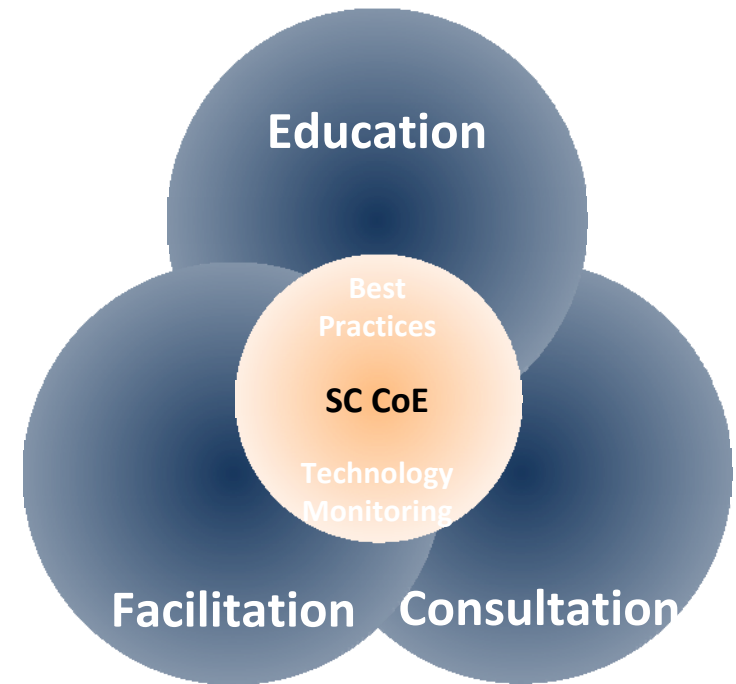
Bio: Ron Daniel, Jr.



- Over 15 years in the business of metadata & automatic classification
 - Disruptive Technology Director, Elsevier
 - Principal, Taxonomy Strategies
 - Standards Architect, Interwoven
 - Senior Information Scientist, Metacode Technologies (acquired by Interwoven, November 2000)
 - Technical Staff Member, Los Alamos National Laboratory
- Metadata and taxonomies community leadership.
 - Chair, PRISM (Publishers Requirements for Industry Standard Metadata) working group
 - Acting chair, XML Linking working group
 - Member, RDF working groups
 - Co-editor, PRISM, XPointer, 3 IETF RFCs, and Dublin Core 1 & 2 reports.

Brought to you by the Smart Content Center of Excellence

- Mission: Support Elsevier in the transition to increasingly more advanced forms of digital publication. Emphasis is helping Product groups see new possibilities.
- The SC CoE will provide:
 - **Education** – Teaching staff and management about Smart Content opportunities, pitfalls, and methods.
 - **Facilitation** – Organize discussions around architecture and the requirements that must shape it. Discussions will include Product, Ops, and IT.
 - **Consulting** - Participate as team members in a few smart content projects.
- The SC CoE will publish and teach best practices for using and creating Smart Content, and
- Helps Elsevier groups anticipate future possibilities by monitoring research and development in the area.



SC CoE Mission

Goals for this talk

- Basic background on metadata, taxonomy, and the terms used in this talk.
- Information on the use of metadata, taxonomies, and other vocabularies
 - In content enhancement
 - In search
 - In content management
- Information on taxonomy selection and management.
 - Tool Use
 - Tool Selection
 - Taxonomy Distribution
- Medium-term applications of ontologies and semi-automated methods for construction.

Pop Quiz

On a blank piece of paper:

- What question(s) did you want to have answered by coming to today's talks?

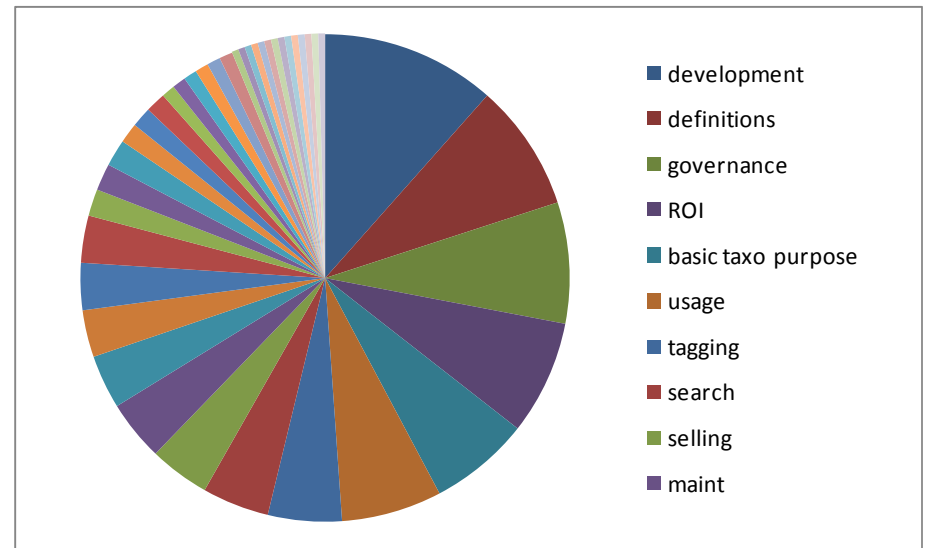
Flag **one** question to be discussed later.

You do NOT have to provide your name.

Please DO provide your job title, division, and either company name or company type.

What do other people ask about?

- How to build a taxonomy?
 - Definitions of terms.
 - How to govern its use and maintenance?
 - What's the ROI?
 - What are they for?
 - How do we put them to use?
 - How do we link them to content?
 - How do they help search?
 - How do I sell management on a taxonomy project?
 - How do we maintain them?
- and many more...*



Agenda

- 9:15 Metadata & Taxonomy Definitions & Background
- 9:30 Use of Metadata and Taxonomy
- 10:00 Use of Taxonomy Tools
- 10:15 Break
- 10:30 Taxonomy Tool Selection
- 11:00 Semi-Automated Ontology Construction
- 11:40 Summary
- 11:45 Questions
- 12:00 Adjourn

Taxonomy and Metadata Definitions

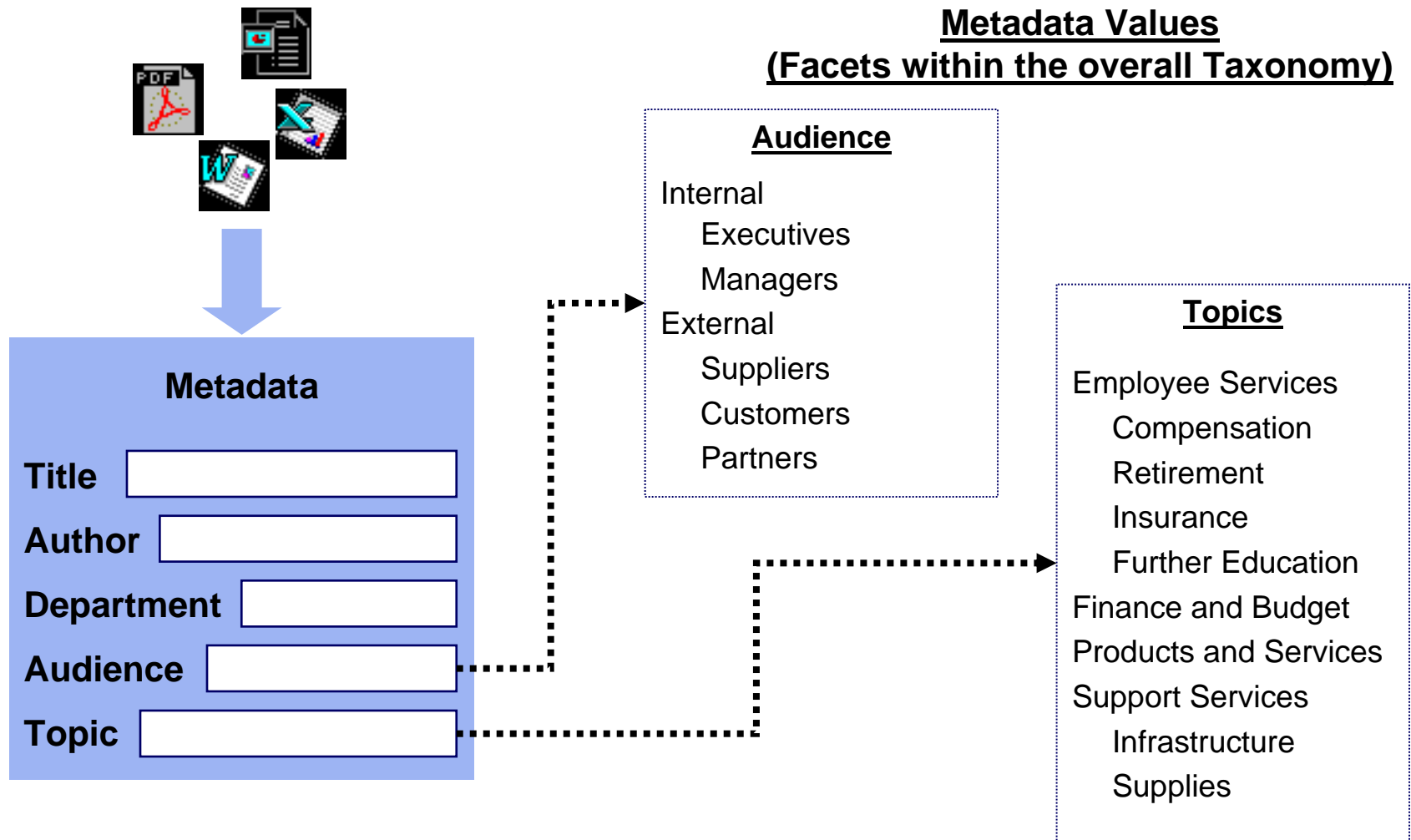
Metadata

- “Data about data”.
- Different communities have very different assumptions about they types of data being described.
 - I’m from the Information Science community, not the database, statistics, or massive storage communities.

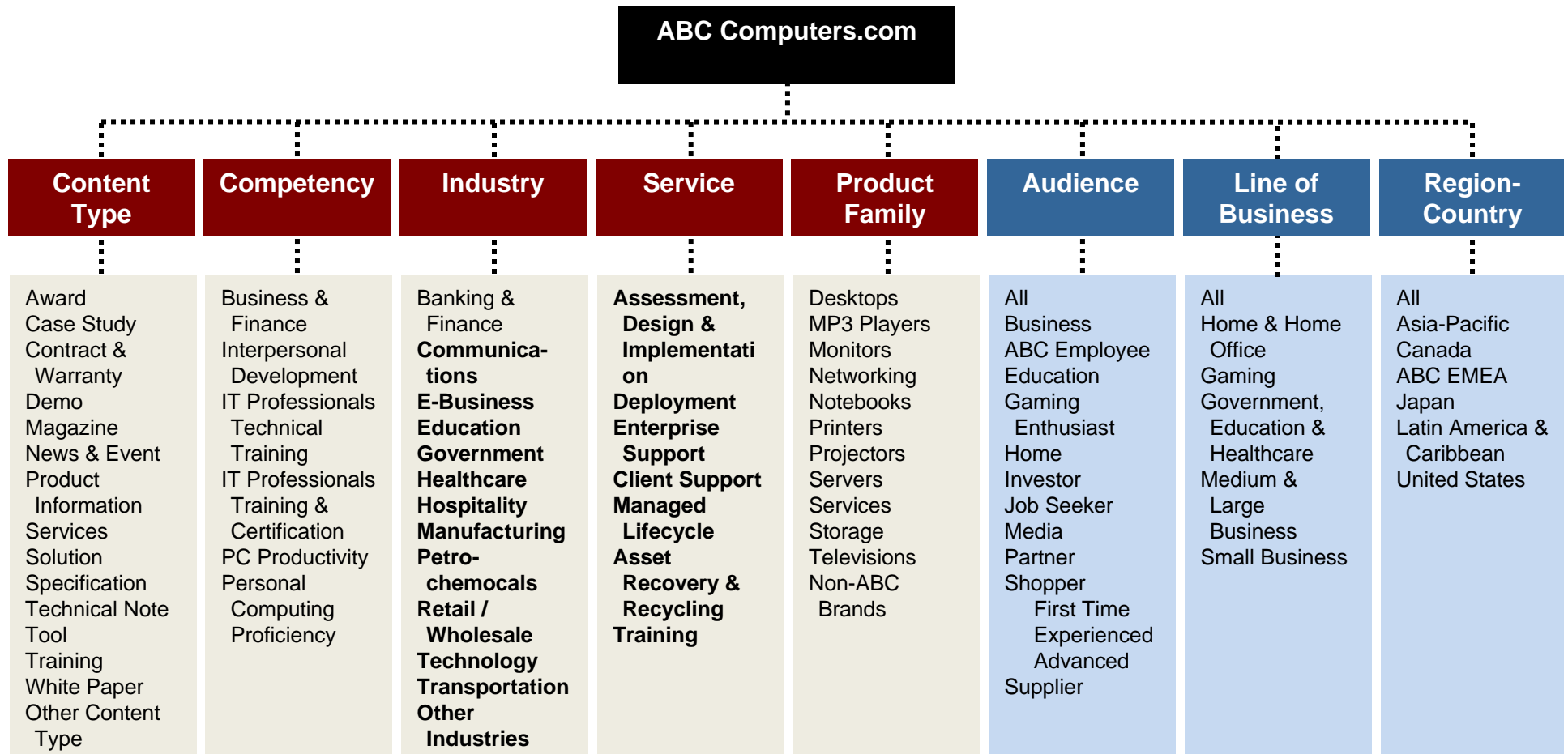
Taxonomy

1. The classification of organisms in an ordered system that indicates natural relationships.
2. The science, laws, or principles of classification; systematics.
3. Division into ordered groups, categories, or hierarchies.

Examples of Taxonomy used to Populate Metadata Fields



Example faceted taxonomy





Manually tagged metadata sample

Attribute	Values
Title	Jupiter's Ring System
URL	http://ringmaster.arc.nasa.gov/jupiter/
Description	Overview of the Jupiter ring system. Many images, animations and references are included for both the scientist and the public.
Content Types	Web Sites; Animations; Images; Reference Sources
Audiences	Educators; Students
Organizations	Ames Research Center
Missions & Projects	Voyager; Galileo; Cassini; Hubble Space Telescope
Locations	Jupiter
Business Functions	Scientific and Technical Information
Disciplines	Planetary and Lunar Science
Time Period	1979-1999

Discussion

- What sorts of facets are you concerned with?

Other kinds of Vocabularies

Type	Remarks
Synonym Ring	<ul style="list-style-type: none"> ▶ Connects a series of terms together ▶ Treats them as equivalent for search purposes e.g (Dog, Canine, Pooch, Mutt) (Cat, Feline, Kitty), ...
Authority File	<ul style="list-style-type: none"> ▶ Used to control variant names with a preferred term ▶ Typically used for names of countries, individuals, organizations e.g. (IBM, Big Blue, International Business Machines Inc.)
Classification Scheme	<ul style="list-style-type: none"> ▶ A hierarchical arrangement of terms ▶ May or may not follow strict “is-a” hierarchy rules ▶ Usually enumerated; ie, LC or Dewey
Thesaurus	<ul style="list-style-type: none"> ▶ Expresses semantic relationships of: <ul style="list-style-type: none"> • Hierarchy (broader & narrower terms) • Equivalence (synonyms) • Associative (related terms) ▶ May include definitions
Ontology	<ul style="list-style-type: none"> ▶ Resembles faceted taxonomy but uses richer semantic relationships among terms and attributes and strict specification rules ▶ A model of reality, allowing inferences to be made.

Agenda

9:15 Metadata & Taxonomy Definitions & Background

9:30 Use of Metadata and Taxonomy

in Content Enhancement

In Search

in Content Management

9:45 Use of Taxonomy Tools

10:15 Break

10:30 Taxonomy Tool Selection

11:00 Semi-Automated Ontology Construction

11:40 Summary

11:45 Questions

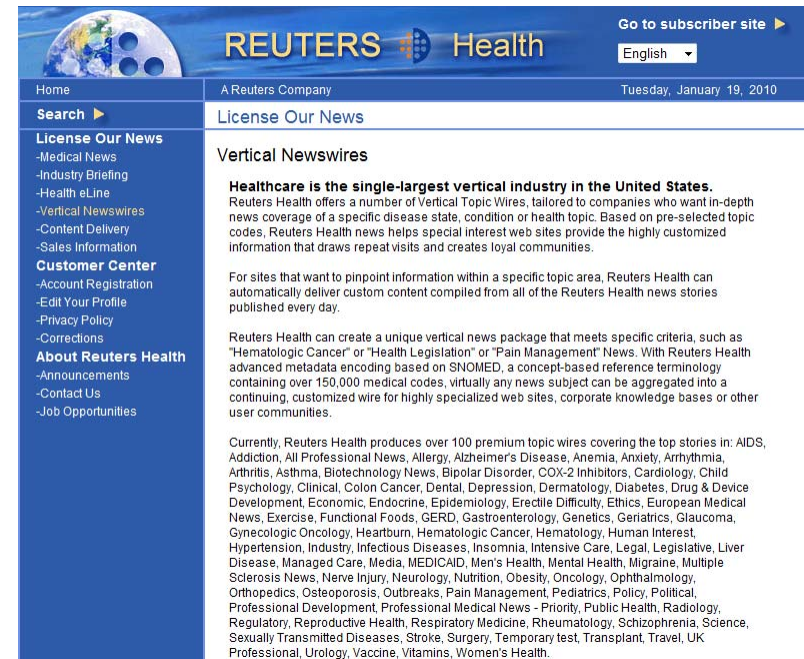
12:00 Adjourn

Case Study: Custom Newswires at Reuters Health (ca. 1999)

- Reuters Health produced two types of medical news stories – professional and consumer.
 - Also produced a small number of topic-based subsets (e.g. AIDS, Breast Cancer, Women’s Health) by editors dragging copies into extra folders.
- Customers wanted *many* more targeted feeds.
 - Editors and Folders process would not scale up.
 - Decided to tag articles with various fine-grained subject codes, then select for the different feeds based on those codes.
- Created multi-faceted taxonomy:
 - Medical Subject (SNOMED), Industry (NAICS), Location (ISO 3166), Drugs & Chemicals (licensed list), Business Topics (custom), etc.
- Updated editorial workflow system to use **semi-automatic classification**
 - Automated suggestion with manual review & correction by writers when submitting, then by editors.
- Created Sales Tool for salespeople to create queries for customers and send them the customized feeds.

Reuters Health: Lessons Learned

- **Still in use 11 years later.**
- Manual correction capability was very important.
 - Automated method alone not accurate enough.
 - Editorial feedback to stop over-tagging by writers.
- Same idea as “Virtual Journals” or personalized RSS feed.
 - Let end-user have their own sales tool.
- Some tagging could be done at story assignment time.
 - Subject of an article or book is known for a long time.
 - Inline tagging must deal with faster changes in topics, companies, etc.



“For sites that want to pinpoint information within a specific topic area, Reuters Health can automatically deliver custom content compiled from all of the Reuters Health news stories published every day.”

“... over 100 premium topic wires covering the top stories in: AIDS, Addiction, ..., Travel, UK Professional, Urology, Vaccine, Vitamins, Women's Health.”

Facet Navigation



Refine Your Selection

Popular Refinements

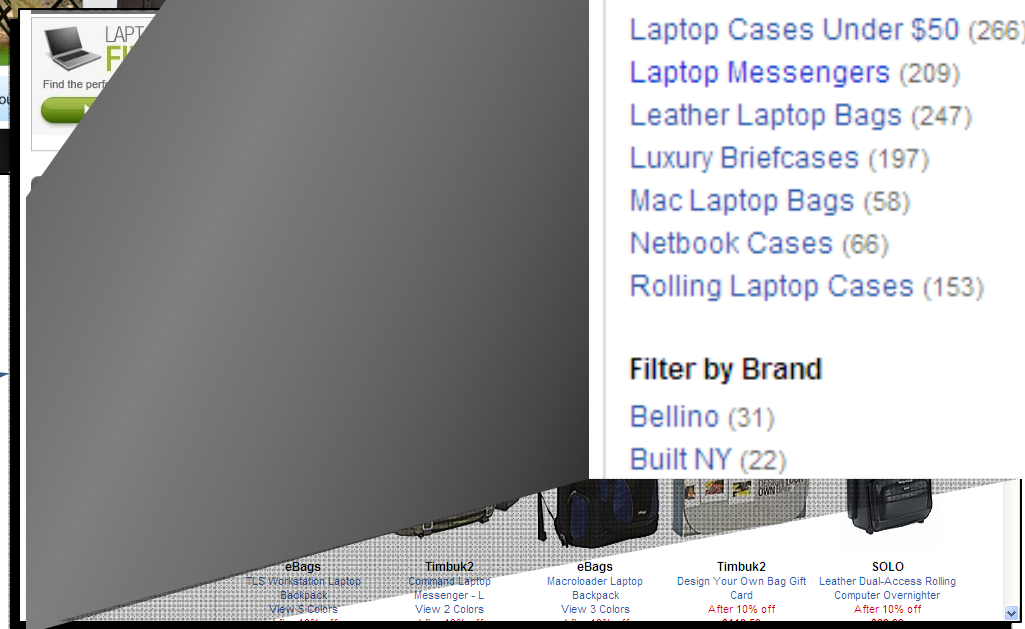
- SALE (595)
- Best of the Best (171)
- New Arrivals (146)

Filter by Category

- Attache (51)
- Checkpoint Friendly L... (98)
- Fabric Laptop Bags (643)
- Laptop Accessories (35)
- Laptop Backpacks (186)
- Laptop Cases For Women (492)
- Laptop Cases Under \$50 (266)
- Laptop Messengers (209)
- Leather Laptop Bags (247)
- Luxury Briefcases (197)
- Mac Laptop Bags (58)
- Netbook Cases (66)
- Rolling Laptop Cases (153)

Filter by Brand

- Bellino (31)
- Built NY (22)



Popular Refinements

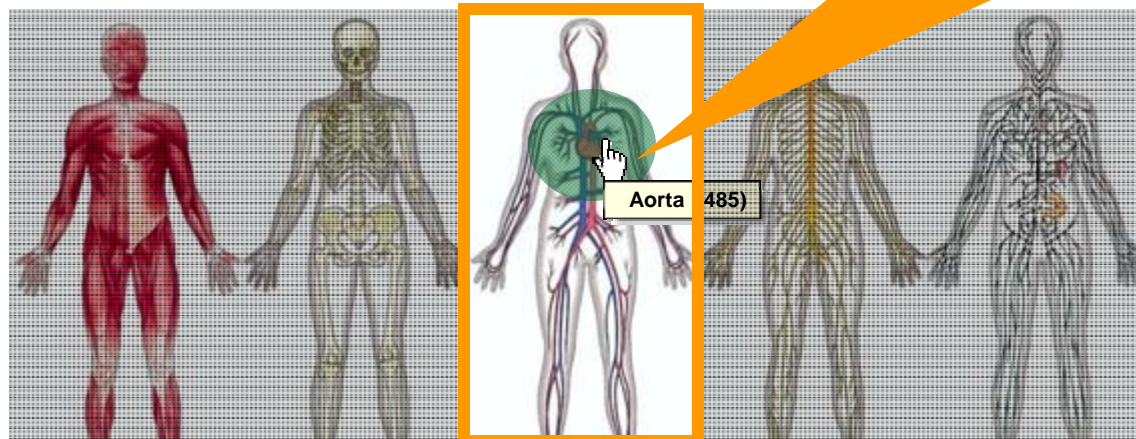
- Filter by Category
- Filter by Brand
- Filter by Color
- Filter by Price
- Filter by Material

Browse by Region and System : Fancy Facet Navigation

Screen Layout Mockup Redacted

- Muscular
- Skeletal
- Vascular
- Nervous
- Lymphatic
- Endocrine
- Digestive
- Reproductive
- ...

Browse by Region and System:



Display multiple systems, with indication of counts by region and popups of more specific areas and counts.

Reuters Health tagging was at the Article level.

What about tagging at a finer level?

ADHD in kids tied to organophosphate pesticides
 Last Updated: 2010-05-17 8:26:21 -0400 (Reuters Health)
 By Frederik Joelving
 NEW YORK (Reuters Health) - Children exposed to pesticides known as organophosphates could have a higher risk of attention-deficit/hyperactivity disorder (ADHD), according to a new study.
 Researchers tracked the pesticides' breakdown products in kids' urine and found those with high levels were almost twice as likely to develop ADHD as those with undetectable levels.
 The findings are based on data from the general US population, meaning that exposure to the pesticides could...

dc:creator	Frederik Joelving
dc:publisher	Reuters Health
dc:title	ADHD in kids tied to organophosphate ...
dc:date	2010-05-17
rhi:medicalSubject	ADHD
rhi:chemicalSubject	Organophosphate
rhi:chemicalSubject	Pesticides
rhi:industrySubject	Pesticide Manufacturers

Entity extraction – Finding the names of people, places, companies, things, dates, events, etc.

Investigation...
 Black Sea...
 on the...
 for downing the aircraft en route from Israel, killing 78 people.

A delegation from Ukraine's defence ministry is due to arrive on Monday in the Russian Black Sea resort of Sochi, where the investigation is centred, following calls on Saturday from Sergei Ivanov, Russian defence minister, for information on live missile fire during Ukrainian military exercises at the time of the crash.

Vladimir Putin, Russian president, was not satisfied with preliminary information supplied by Ukraine, according to Mr Ivanov, who said Alexander Kuzmuk, Mr Ivanov's deputy, was not sufficiently convinced.

The comments...
 Russia is prepared...
 was involved. P...
 details of such...

People	Sergei Ivanov; Vladimir Putin; Alexander Kuzmuk; Mr Ivanov
Countries	Russia; Ukraine; Israel
Towns	Sochi
Geographic Features	Black Sea
Organizations	Siberia Airlines
Events	crash
Objects	Tu-154
Dates	Sunday; Monday; Saturday
Quantities	78

Information Extraction (IE)

- Recognizing facts based on patterns of extracted Entities
 - Compliance Monitoring Problem: Find illegal disease benefit claims for companies selling natural supplements on the web.
 - <Substance> <Claim> <Disease>
 - Ginseng helps with Diabetes
 - Competitive Intelligence Problem: Monitor personnel movements in an industry.
 - <Person> <Role> <Organization>
 - John Smith, CEO of XYZ Corp



Triples can be pulled out of large amounts of text and organized for review and action.

Image courtesy of Lingustat

Agenda

9:15 Metadata & Taxonomy Definitions & Background

9:30 Use of Metadata and Taxonomy

9:45 Use of Taxonomy Tools

10:15 Break

10:30 Taxonomy Tool Selection

11:00 Semi-Automated Ontology Construction

11:40 Summary

11:45 Questions

12:00 Adjourn

Term management functional requirements

Basic

- Standard and Custom Fields
- Standard and Custom Relations
- Data Typing and Restrictions
- Consistency Enforcement
- Polyhierarchy
- Term Search
- Flexible Reporting

Basis for selection by Vocab name

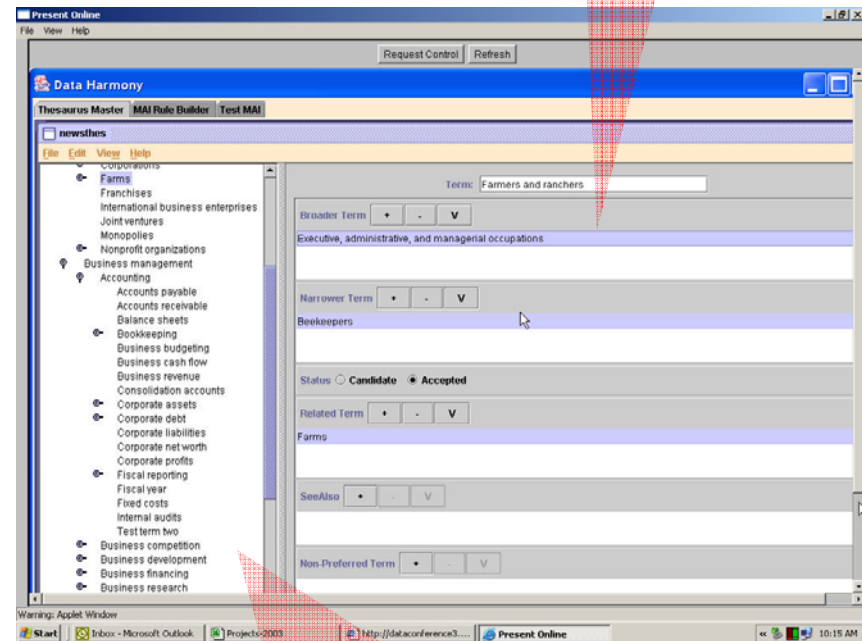
Midrange

- UNICODE
- Unique IDs
- Multiple Vocabulary Support
- Inter-Vocabulary Mapping
- Specifiable Term Ordering
- Audit Trail
- Multi-User Security

Entity Editing

Advanced

- Persistent IDs (Namespaces)
- Merge & Unmerge Multiple Vocabularies
- Business Rules Programmability
- Editorial Rules Enforcement
- Change Request Workflow
- Voting



Hierarchy Browser

Term management functional requirements: Details

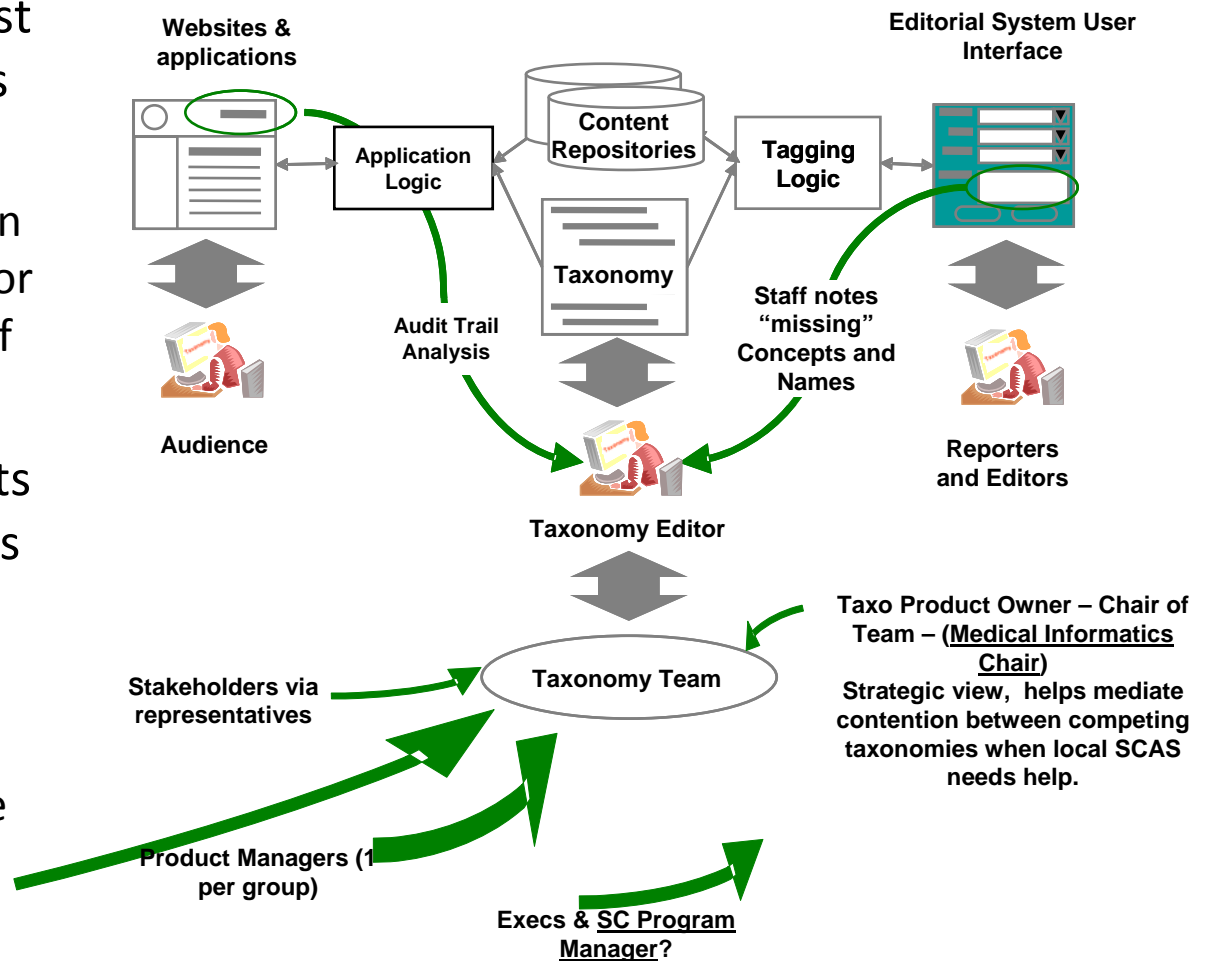
- Aliases – Support synonyms, quasi-synonyms as well as alternative labels based on language or other factors.
- Notes – Multiple types of notes fields, e.g., to keep public notes separate from editorial working notes.
- Effective dates? – View a ‘valid’ taxonomy state as it was at any point in time.
- Merge & Unmerge – Create a union of two or more records, and be able to undo it?
- Manage multiple versions of external taxonomy and the ‘official’ crosswalk provided.
- Inter-category relations – Provide links between vocabularies that may or may not have the same hierarchy. Build views using different hierarchies
- Poly-hierarchy – Basic tools handle a term and all its children. Mid-range tools should handle a term with or without children.
- Rules enforcement – Check conformance to style rules like length, use of & vs. “and”, etc.
- Workflow – Track change requests, facilitate approval process, and report on status.

Additional Functional Requirements

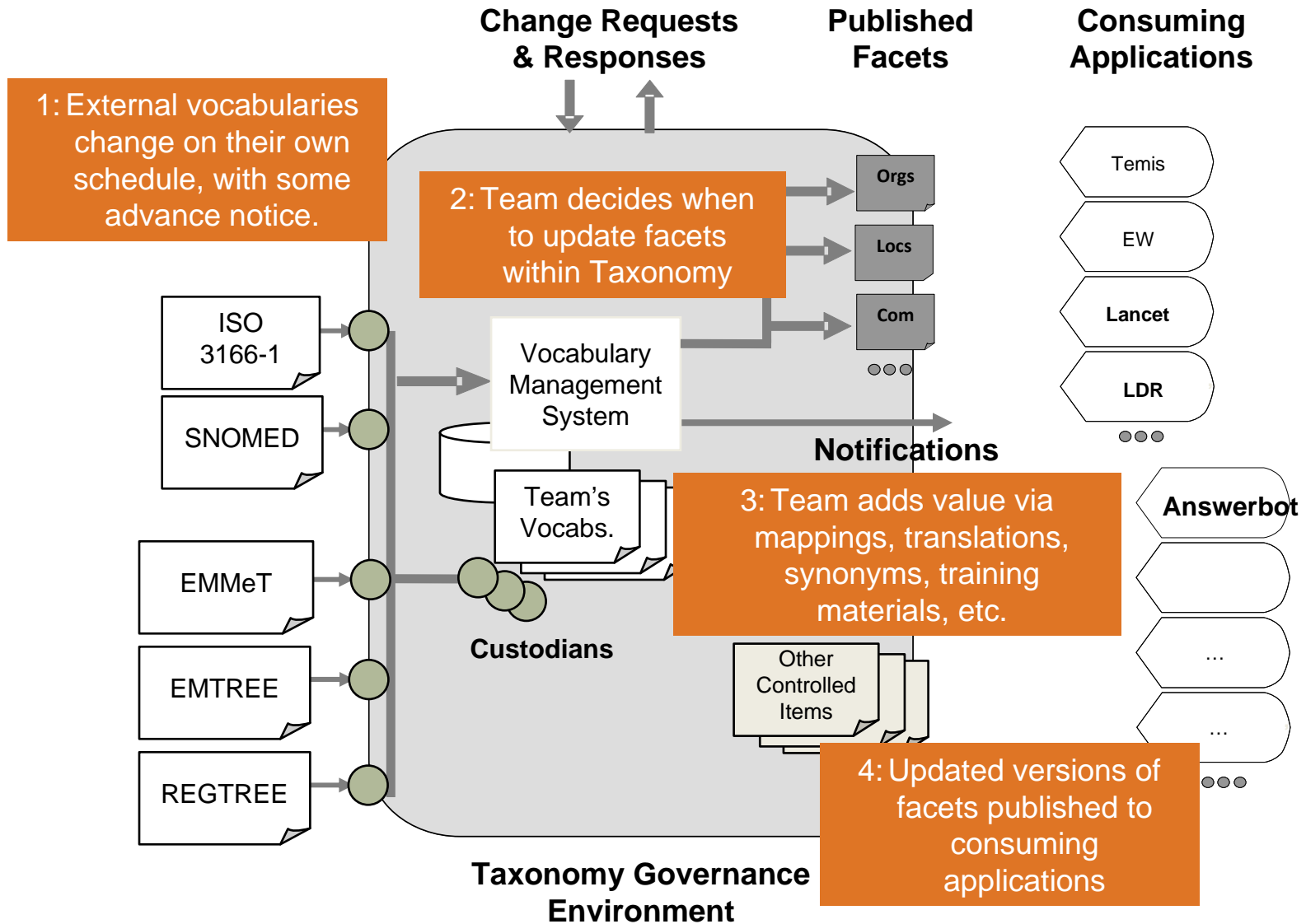
- Requirements that are commonly missed:
 - Capacity
 - How many terms (entries & variants) can be supported in one taxonomy?
 - How many taxonomies can be supported in one application?
 - Performance impacts vs. taxonomy size
 - Hardware requirements for acceptable performance
 - Price, TCO, License terms and conditions
 - *Specifics* of integration with *specific* other tools
 - e.g. “Can your tool read and write format X out-of-the-box? If not, what will be the price to develop a converter so it can?”
- **The biggest point: Define the processes, in at least moderate detail, before procuring a tool.**

Tagging and Taxonomy Workflow

- Reporters and Editors must not do taxonomy or Temis changes on the fly.
- Their role is to note errors in tagging and mark the text for easy automated insertion of corrections once approved.
- Taxonomy change requests come from several sources and are considered by a Taxonomy Team.
- Once the team decides, the Taxonomy Editor makes the changes.



Overview of typical governance environment



Selection process

- Ontology tool selection can use a typical selection process:
 - Define use cases, Infer requirements, Weight criteria, Ask vendors, Score results
 - Criteria should include ease of use, ease of integration, cost, version control, schema design flexibility, and auto-analysis capabilities.
- Checking technical IT “gotchas” is good, but get at the business process first. Use cases *must*:
 - Start from a definition of the business processes to be instituted
 - Get into details of how taxonomies will be created, used, and maintained.
- Otherwise you end up with overly-general requirements and no motivation for them:
 - e.g. “Can your tool export selected parts of the taxonomy and ontology”?

Tool and Process Integration *Matters!*

- How will downstream software (e.g. tagging tools, search and navigation tools) deal with taxonomy changes?
- What are characteristics of various vocabularies (size, need for inter-relationships, volatility, etc.).
- How will editors ask for new tags, or indicate that a term is a synonym of an existing tag? How will those requests be handled?
- How will reader detection and staff correction of tagging errors be handled?
- How will tracking of correct and incorrect tags for continual improvement be handled?
- What are data volumes, data rates, response time requirements, etc.?
- How will taxonomy information affect search or other applications?

Potential missing requirements

- Some means of getting and tracking change requests is needed.
 - Could be in the tool, OR
 - Could be a simple external bug-tracking database.
 - Who needs to be informed about changes? Do we need a RACI model?
- Read and Request access for Project Managers vs. general staff?
- What output formats and methods are required?
 - CSV? .XLS? SQL? WSDL? SPARQL? SKOS? ZTHES?
 - Which systems are communicating and what information do they need to exchange?
- What access control is needed? Do we need to limit access to different

Key Standards re. Taxonomy Management

- Unicode, XML, xml:lang
- RDF, RDFS, OWL*
- ISO 5496 –
 - Guidelines for the Establishment and Development of Multilingual Thesauri
 - Others: Z39.19
- SKOS, SKOS-XL
 - SKOS was developed to model Concepts in the world, not the names of concepts. SKOS-XL helps fix that. Big help with multilingual.
- UMLS Semantic Relations

UMLS Semantic Relations

isa
associated_with
physically_related_to
part_of
consists_of
contains
connected_to
interconnects
branch_of
tributary_of
spatially_related_to
location_of
adjacent_to
surrounds
traverses
...

Agenda

9:15 Metadata & Taxonomy Definitions & Background

9:30 Use of Metadata and Taxonomy

9:45 Use of Taxonomy Tools

10:15 Break

10:30 Taxonomy Tool Selection

11:00 Semi-Automated Ontology Construction

11:40 Summary

11:45 Questions

12:00 Adjourn

Agenda

- 9:15 Metadata & Taxonomy Definitions & Background
- 9:30 Use of Metadata and Taxonomy
- 9:45 Use of Taxonomy Tools
- 10:15 Break
- 10:30 Taxonomy Tool Selection
- 11:00 Semi-Automated Ontology Construction
- 11:40 Summary
- 11:45 Questions
- 12:00 Adjourn

Fun Questions

The animals are divided into:

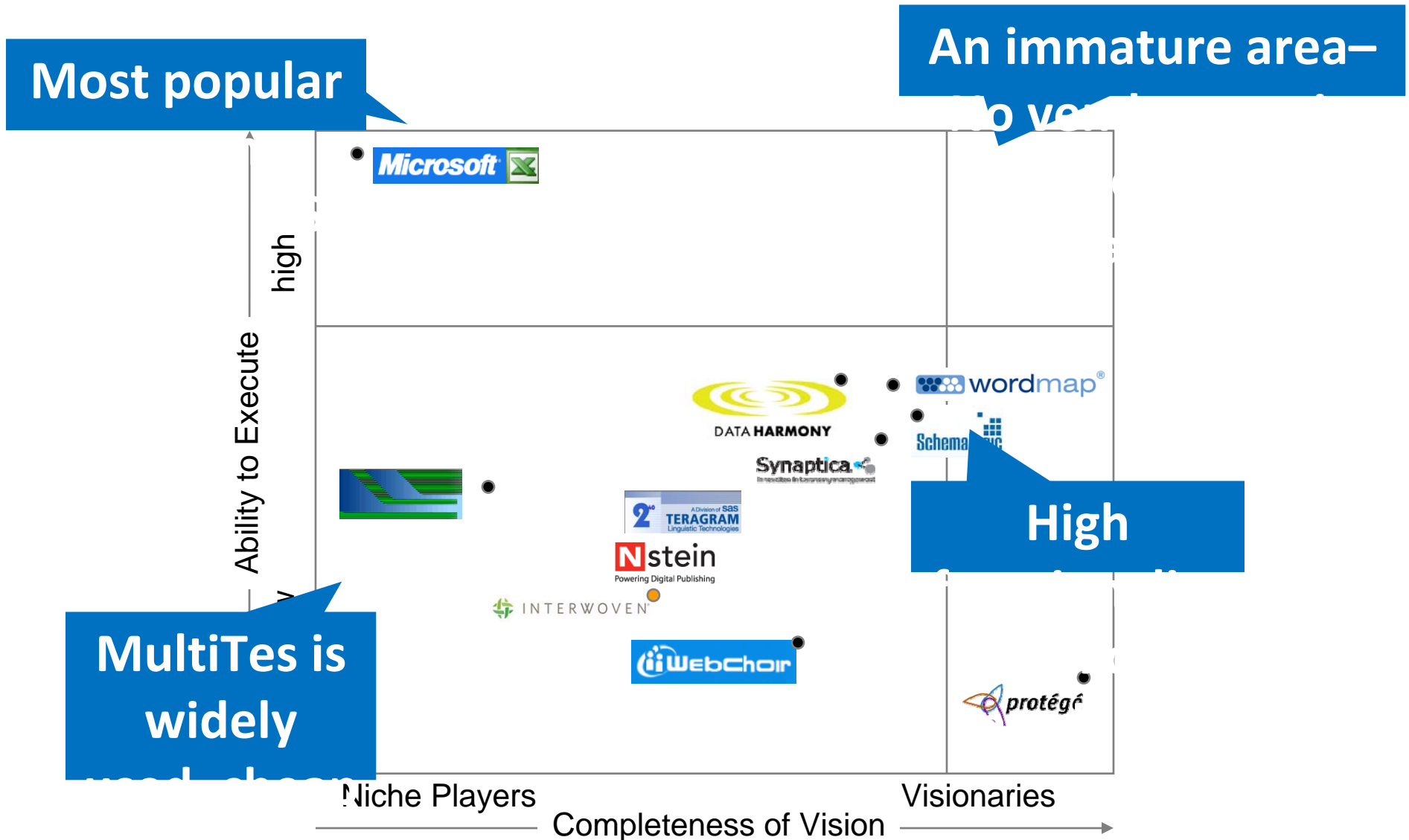
(a) belonging to the emperor, (b) embalmed, (c) tame, (d) sucking pigs, (e) sirens, (f) fabulous, (g) stray dogs, (h) included in the present classification, (i) frenzied, (j) innumerable, (k) drawn with a very fine camelhair brush, (l) et cetera, (m) having just broken the water pitcher, (n) that from along way off look like flies.

This was created to be as bad a classification as possible. What makes it so bad?

Jorge Luis Borges, " THE ANALYTICAL LANGUAGE OF JOHN WILKINS"

Works in 3 volumes (in Russian). St. Petersburg, "Polaris", 1994. V. 2: 87.

Ranking taxonomy editing tools & vendors – according to Taxonomy Strategies



Timing of Ontology Tool Selection

- Full Ontology Tool process will take significant amount of time:
 1. Defining the business process for creation and maintenance,
 2. Conducting a selection process,
 3. Procuring the tool,
 4. Installing and configuring it*,
 5. Loading it with values from different sources,
 6. Harmonizing the overlaps,
 7. Training and operationalizing the tool.
- Any of these can be short-circuited, but the result will be more difficult and expensive maintenance, and a higher probability of errors.
 - Excel is perfectly appropriate for sketches and early prototypes
 - Excel is NOT appropriate for maintenance
- Customization is very common in taxonomy editing tools, so implementation and configuration can be expensive.
 - Input and output formats, specific fields for specific vocabularies, specific relationships between vocabularies, etc.



Questions re. Workflow Requirements

- Does tool have a built-in status model for changes?
 - If so, does it fit or can it be modified or side-stepped?
 - If not, can it be added easily and still leave room for other customized fields?
- What status codes are needed for the workflow?
 - MultiTes provides Candidate, Provision, Approved, Not Valid.
- Tool does not HAVE to enforce a workflow, but it SHOULD provide basic elements of process:
 - A status field, approval date, removal date, effective dates, etc.
 - Workflow may be enforced by other tools.

The screenshot shows a software window titled "Western Africa" with a menu bar (Clipboard, Edit, Print, Window) and tabs (Record Details, Multilevel Hierarchy, 2-way Hierarchy, Edit). The form contains the following fields and values:

Term	Western Africa	Save
Approval date	2009-04-14	Cancel
Not valid date		
Status	Approved	
Flag	Approved	
TNR	Candidate	
Input date	2009-04-14	By JBusch
Last Updated	2009-04-14 10:57	By JBusch

Buttons: Close All, Close, Stay on top (checked), ENG

Sample Selection Plan

Actions	O	O	O	O	N	N	N	N	N	D	D	D	D	J	J	J	J	J	F	F	F	F	M	M	M	M	
	4	11	18	25	1	8	15	22	29	6	13	20	27	3	10	17	24	31	7	14	21	28	7	14	21	28	
Team Formation	█	█	█																								
Rough Plan, SharePoint		█	█																								
Gather and Evaluate Previous Materials	█	█	█	█	█	█	█	█	█	█	█	█	█														
Create Full Use Case document				█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█	█
Create Full Requirements List	█	█	█	█	█	█	█	█	█	█	█	█	█														
Create Long Vendor List				█	█	█	█	█	█	█	█	█	█														
Create Basic Requirements List				█	█	█	█	█	█	█	█	█	█														
Early Knockout Rounds								█	█	█	█	█	█														
Create Medium Vendor List														█	█	█	█	█	█	█	█	█	█	█	█	█	█
Develop RFP														█	█	█	█	█	█	█	█	█	█	█	█	█	█
Gather and Evaluate Responses																											
Invitational webinars																											
Select final candidates (n=?)																											
Test selected taxonomy tools in-house																											
Develop recommendation																											
Negotiations																											

Basic Requirements: Scale, Multiuser, Large Updates, Multilingual, Environment

1. What is the largest vocabulary size, both in number of terms and file size, you know your system has successfully loaded?
2. How many different users can edit one vocabulary at the same time? If one user saves a change, is it immediately visible to the other users?
3. Does your system allow the import of bulk data into an *existing* taxonomy?
4. Does your system support UNICODE? Do you have examples of multi-lingual vocabularies built and maintained in the system?
5. Does your system run on the Windows 64-bit platform? (in 64 bit mode)

Vendor	Software
ACS 121	One 2 One
Altova	Altova
Apelon	Terminology Management
Applied Relevance	AR.Taxonomy
Arity	LexiLink
Cuadra	STAR/Thesaurus
Data Harmony	Thesaurus Master
DOME	DERI Ontology Management Environment
Hozo	Hozo Ontology Editor
Interwoven	MetaTagger
Microsoft	Excel
Mondeca	Intelligent Topic Manager (ITM)
MultiTes	MultiTes Pro
Ontopia	Ontopia
Open Text	Taxonomy Manager
Pool Party	Pool Party
Protégé	Protégé
Revelytix	knoodl.com
SAS	SAS Ontology Management
SchemaLogic	Schema Server
Smartlogic	Semaphore Ontology Management
Soutron	SoutronTHESAURUS
Synaptica	Synaptica
TemaTres	TemaTres Vocabulary Server
Thesaurus Builder	Thesaurus Builder
Tim Craven	TheW32
TopQuadrant	TopBraid suite
University of Zaragoza / GeoSpatiumLab	ThManager
WAND	Webchoir
Wordmap	Wordmap Designer

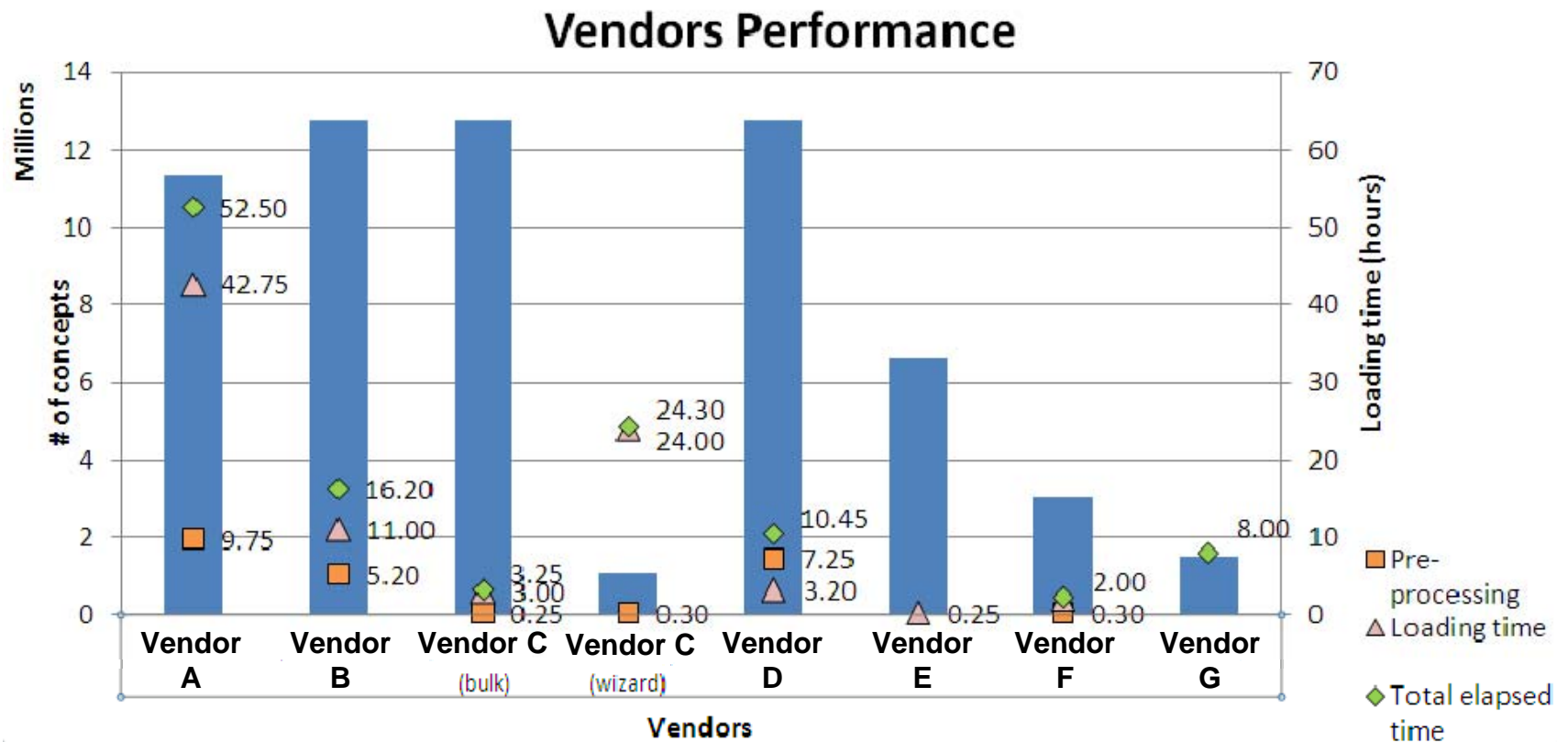
Knockout Round #2 - Scale

- Instructions:
 - Load TrEMBL (18M concepts, VERY simple structure)
 - Copies are at this URL
- Getting the usual questions...
 - “Can I get this in a different format”?
 - “Can I use this other file instead ”?
 - “Do I really have to load all of it ”?
 - “Can I get documentation on that file first”?
 - “SKOS is not good at protein data, we can modify that”.
 - “Oh, that won’t take long at all”.
 - “This can’t really be representative!”

**Portions
of this
slide have
been
redacted**

Tool Scalability Test

- Blue Columns are number of items loaded
- Datapoints are time it took
- Cut from 15 to 4-ish



Use Cases for Deriving More Detailed Requirements

3.1 Selecting Terms for Entity Matching Lexicons

3.2 Lightweight Vocabularies

3.3 Utility Vocabularies

3.4 Vocabulary Discovery

3.5 Quality Control Checking for Vocabulary Importing

3.6 Vocabulary X requirements.

3.7 Mapping to Linked Data Hubs

3.8 Vocabulary Suitability Testing

3.9 Continuation of Vocabulary A,B,C... Maintenance

3.10 Merged Vocabulary Construction and Validation

**Portions
of this
slide have
been
redacted**

Current Status

- Vendor Webinars underway now
- Created testing plan for more in-depth work with our data, typically on a hosted instance.
- Recommendation due in a few weeks.

Agenda

- 9:15 Metadata & Taxonomy Definitions & Background
- 9:30 Use of Metadata and Taxonomy
- 9:45 Use of Taxonomy Tools
- 10:15 Break
- 10:30 Taxonomy Tool Selection
- 11:00 Semi-Automated Ontology Construction
- 11:40 Summary
- 11:45 Questions
- 12:00 Adjourn

Fact Extraction

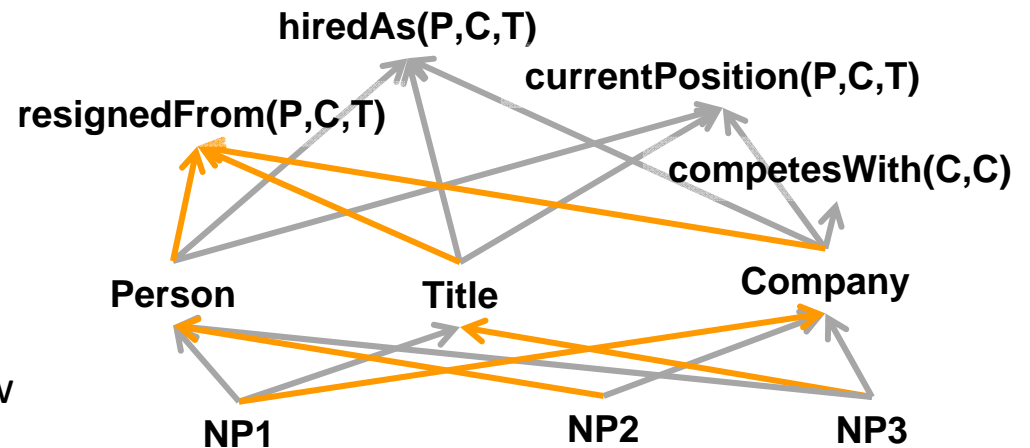
- Fact Extraction builds on results of Entity Extraction, and uses Patterns of connections between Entities.
 - e.g. a competitive intelligence application might look for people who are changing jobs:
 - PERSON “, the new” JOBTITLE “of” ORG
 - ORG “announced that” PERSON “has been hired as” JOBTITLE
 - PERSON “retired from” ORG
 - etc.
 - Different applications look for different types of entities and different patterns
 - Clinical trials, Illegal claims, substances affecting organs, gene-protein-expression, ...
- We want the computer to learn new Entities and Patterns:
 - New entities might be new people, diseases, drugs, products, events, etc.
 - New patterns will increase the number of facts that can be found (improved recall)

How are Facts Extracted?

- Both Rule-Based and Learned Approaches exist.
- What about Accuracy?
 - Entity Recognition, Part of Speech tagging, Fact Extraction, and Learning all have their own error rates.
 - Combined error rates could be terrible!
- Solution – add more constraints.
 - Possible new entity or pattern must work consistently as part of many different “facts” before it is believed and added to the list of facts.

Learning More Facts

- Manually build:
 - A Model of how this part of the world works.
 - Initial Vocabularies of names of People, Places, and other things. (Could be pre-existing lists).
 - Some sentence Patterns that show the facts wanted.
- Computer loops:
 - Add high-confidence entities
 - Add high-confidence patterns
 - Extract Facts
 - Mark potential entities
 - Mark potential patterns



Sears announced the resignation of John Smith as their CEO, ...

Sears announced CEO, John Smith, ...

Sears CEO, John Smith, said ...

... according to John Smith, CEO of Sears.

John Smith resigned as CEO of Sears ...

Sample of Learned Patterns for Companies

- Knowing a few high-confidence entities and patterns lets us learn facts involving those entities.
- Using partially-completed patterns lets us learn new entities and patterns.
- If those new entities and patterns appear several times, we can add them to the list of known entities and patterns, then learn new facts from them.

advertisers like C,
advertisers such as C,
chains like C,
chains such as C,
competitors like C,
a company like C,
a big company like C,
companies like C,
companies including C,
corporations like C,
discounters like C,
firms like C,
retailers like C,
stores like C,
an operating business of C,
being acquired by C,
a senior manager at C,
a licensing deal with C,
an executive at C,
a software engineer at C,
...

Tom Mitchell, "How will we Populate the Semantic Web on a Vast Scale?" Keynote at 2009 International Semantic Web Conference, http://rtw.ml.cmu.edu/slides/RTW_ISWC_mod_Oct2009.pdf

Can Taxonomies be built Automatically?

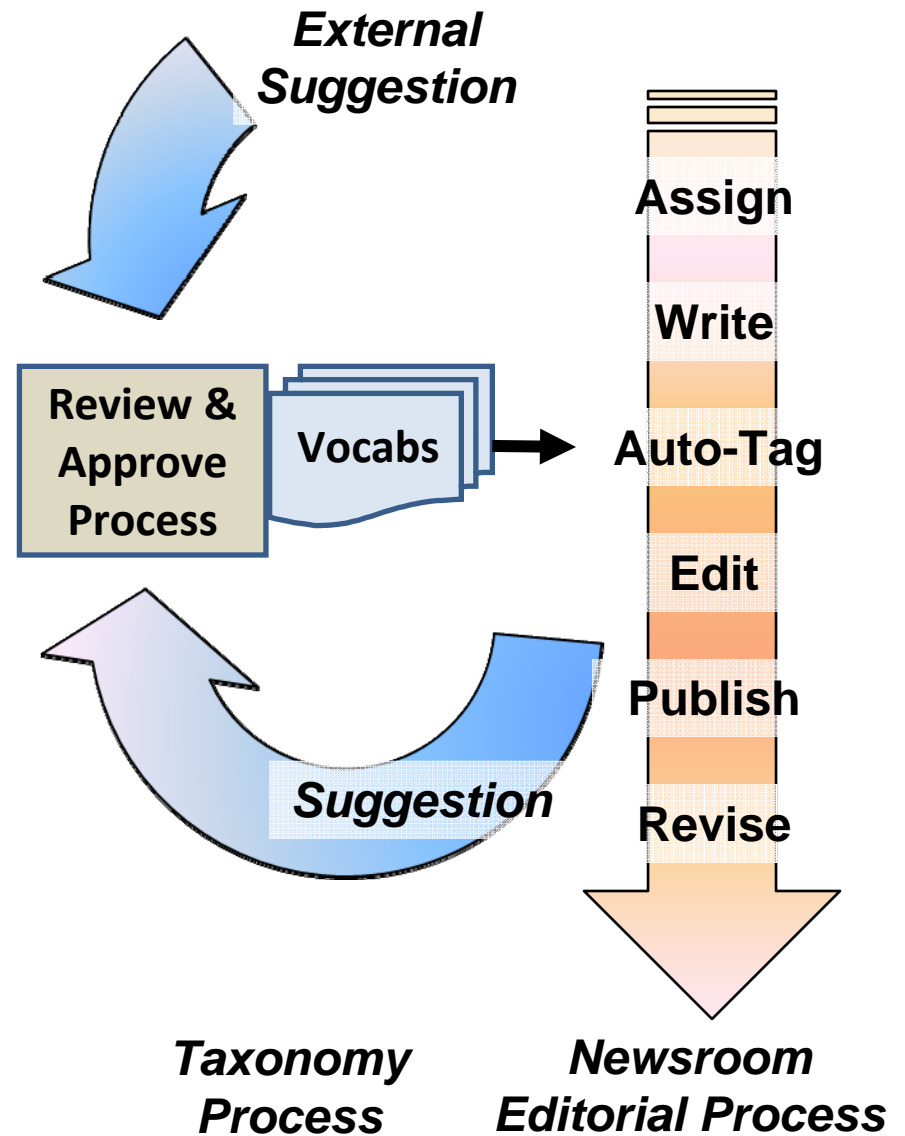
- Software can scan large quantities of content and extract statistically significant words and phrases.
- Example: Archive of 10 publications was analyzed for topics significant to 'copyright'.
- Software does a poor job of
 - de-duplication
 - turning those significant words and phrases into a larger structure
 - discriminating between gold and garbage
- Software is good for
 - getting an understanding of the key phrases in a large amount of content
 - providing test cases for evaluating a taxonomy



Source: Sample data courtesy of Randy Marcinko and nStein.

Best Practice is Hybrid Approach for Maintenance

- Editorial and Taxonomy Processes must interact.
- Editorial-stage corrections to tagging:
 - Discovers new terms
 - Discovers synonyms
 - Discovers homographs
 - Improves categorizer accuracy through better training sets.
- Improved tagging reduces editorial burden.



Agenda

- 9:15 Metadata & Taxonomy Definitions & Background
- 9:30 Use of Metadata and Taxonomy
- 9:45 Use of Taxonomy Tools
- 10:15 Break
- 10:30 Taxonomy Tool Selection
- 11:00 Semi-Automated Ontology Construction
- 11:40 Summary
- 11:45 Questions
- 12:00 Adjourn

Summary

- Vocabularies of various types are key to effective information management.
- Variety of tools exist; there are many different sets of starting assumptions.
- If an organization manages ONE vital vocabulary, a full-custom system that evolves over time is typical.
- Organizations that must manage multiple vocabularies will find a tool helpful.
- Be prepared for significant configuration and customization effort – vocabularies are surprisingly different in structure and use and must be maintained in different ways.

Agenda

- 9:15 Metadata & Taxonomy Definitions & Background
- 9:30 Use of Metadata and Taxonomy
- 9:45 Use of Taxonomy Tools
- 10:15 Break
- 10:30 Taxonomy Tool Selection
- 11:00 Semi-Automated Ontology Construction
- 11:40 Summary
- 11:45 Questions
- 12:00 Adjourn

Agenda

- 9:15 Metadata & Taxonomy Definitions & Background
- 9:30 Use of Metadata and Taxonomy
- 9:45 Use of Taxonomy Tools
- 10:15 Break
- 10:30 Taxonomy Tool Selection
- 11:00 Semi-Automated Ontology Construction
- 11:40 Summary
- 11:45 Questions
- 12:00 Adjourn



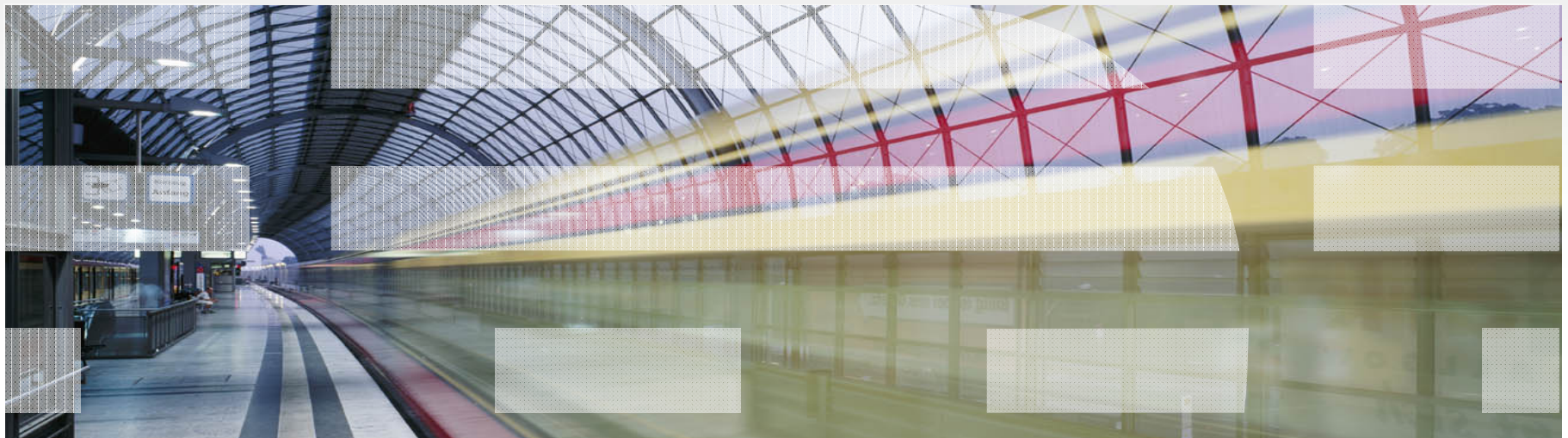
ELSEVIER

Contact Info

Dr. Ron Daniel, Jr.

+1 619 208 3064

r.daniel ~at~ elsevier.com



ELSEVIER

Building Insights. Breaking Boundaries.™