Metadata and Taxonomy
Taxonomy Strategies

- Business consultants who specialize in applying taxonomies, metadata, automatic classification, and other information retrieval technologies to the needs of business and government.
- Leadership in enterprise content management, knowledge management e-commerce, e-learning and web publishing.
- Spin-off from Metacode Technologies, developer of XML metadata repository, automated categorization methods and taxonomy editor acquired by Interwoven in 2000 (now part of Autonomy).
- More than 30 years experience in digital text and image management.
- Metadata and taxonomy community leadership.
  - President, American Society for Information Science & Technology
  - Dublin Core Metadata Initiative Board Member
  - American Library Association Committee on Accreditation External Reviewer

Recent taxonomy projects

http://www.taxonomystrategies.com/html/clients.htm
What do you hope to get out of this workshop?
Agenda

- Interoperability
- Named entities exercise
- Problems with metadata
- Two types of vocabularies
- Dates, roles and topics
- Types of semantic schemes
- Tagging assets
- Tools for automating tagging
Interoperability

- The ability of diverse systems and organizations to work together by exchanging information.
- Semantic interoperability is the ability to automatically interpret the information exchanged meaningfully and accurately.
Interoperability ROI

- Assets are expensive to create so it’s critical that they can be found, so they can be used and re-used.
- Every re-use decreases the asset creation cost and increases the asset value.
Interoperability (2)

❖ If assets are so important, why can’t they be found?
  ▪ They contain no searchable text.
  ▪ They exist in different applications, file shares and/or desktops.
  ▪ … Other reasons?

❖ When they are found why can’t assets be reused?
  ▪ When there are multiple versions, it’s difficult to choose which one to use.
  ▪ The usage rights may not be clear.
  ▪ … Other reasons?
**Interoperability (3)**

- Digital assets are sourced from multiple applications and locations
  - Product lifecycle management (PLM) application
  - Product information management (PIM) application
  - Third party contractors’ systems
  - In-house graphic design department
  - Marketing and Communications servers
  - …Other applications and locations?
Interoperability vision

- I want to easily find any assets in a particular format that can be used for a specific purpose regardless of where they are located.

- Challenges:
  - How to align different metadata properties
    - E.g., Title and Caption; Location and Setting; etc.
  - How to align different vocabularies
    - E.g., CA and California; RiM and Research in Motion; etc.
Agenda

- Interoperability
- Named entities exercise
- Problems with metadata
- Two types of vocabularies
- Dates, roles and topics
- Types of semantic schemes
- Tagging assets
- Tools for automating tagging
People

Arroyo, Gloria
Atkinson, Rowan
Baldwin, Alec
Bale, Christian
Banderas, Antonio
Baron Cohen, Sacha
Barrymore, Drew
Beck, Glenn
Brand, David
Bush, George
Berry, Halle
Kissinger, Henry
Bin Laden, Osama
Blair, Tony
Blanchett, Cate
Bono
Branson, Richard
Brown, Dan
Brownback, Sam
Bryant, Kobe
Buffett, Warren
Burton, Tim

Gingrich, Newt
Giuliani, Rudolph
Goldberg, Whoopi
Gore, Al
Grace, Nancy
Grant, Hugh
Griffin, Kathy
Orisham, John
Hagel, Chuck
Hondros, Tim
Hanks, Tom
Hearn, Sam
Hart, John
Hastert, Dennis
Hawkeye
Hawking, Stephen

Powell, Colin
Prince
Prince Charles
Prince William
Putin, Vladimir
Qaddafi, Muammar
Queen Elizabeth II
Queen Latifah
Rather, Dan
Reid, Tara
Reynolds, Bud
Richardson, Bill
Rivera, Geraldo
Robbins, Tim
Roberts, Julia
Robertson, Pat
Rock, Chris
Rove, Karl
Rowling, J.K.
Sandler, Adam
Santorum, Rick

Who are some important people whose names should be managed? … and why? …

* courtesy of mondostars.com
What are some important organizations whose names you need to manage? ... and why? ...
What are some important products and services whose names you need to manage? ... and why? ...
What are some key events whose names you need to manage? ... and why? ...
What are some significant locations whose names you need to manage? ... and why? ...
What are managed vocabularies

- Names of people, organizations, products, events, locations, etc.
  - Alternate labels
    - Synonyms
    - Abbreviations
    - Acronyms
    - etc.
  - Additional information
    - Unique identifiers
    - Coverage dates
    - Descriptions
    - etc.

- A set of concepts, optionally including statements about semantic relationships between those concepts.
Agenda

- Interoperability
- Named entities exercise
- Problems with metadata
- Two types of vocabularies
- Dates, roles and topics
- Types of semantic schemes
- Tagging assets
- Tools for automating tagging
Problems with metadata

- Inconsistent category assignments
  - CA vs. California
  - RiM vs. Research in Motion

- Changes to classification systems over time
  - ICD-9 vs. ICD-10
  - SIC vs. NAICS

- Use of multiple overlapping or different categorization schemes
  - States vs. SMSA’s
  - ICD-9 vs. CDC Diseases and Conditions
  - NASA Taxonomy vs. NASA Thesaurus
Case Study: Inconsistent categories (1)

Problem:

- Inaccurate reporting with incorrect product counts at global health and beauty products company.
- Some SKUs are sold as units, as well as a part of a kit, a set and/or a bill of materials.
- Lacked a consistent, standard language to enable data sharing including:
  - Rules for SKUs.
  - Business processes related to product data.
  - Product data definitions.
  - Single owner for data elements.
  - Roles and responsibilities related to product data.
  - Product data integration points and relationships.
Case Study: Inconsistent categories (2)

**Solution:**

- Faceted SKU taxonomy instead of a single, monolithic taxonomy tree
  - More flexible design.
  - Describe every item with a combination of facets.
  - Focus on *universal facets* applied to all products, or to all products within a large grouping such as a product line.
Case Study: Multiple categorization schemes (1)

Problem:
- Need to promote agency *behavioral health* program to heterogeneous audiences:
  - Human services professionals
  - Concerned family
  - Policy makers
- Merge heterogeneous information sources:
  - Alcohol and drug information
  - Mental health information
  - Other agency and inter-agency resources
    - Drug Abuse Warning Network (DAWN)
    - Treatment Episode Data Set (TEDS)
    - Uniform Reporting System (URS)
Case Study: Multiple categorization schemes (2)

Solution:

- Faceted content tagging and navigation taxonomy
  - Powers the SAMHSA Store as illustrated in a YouTube video
  - The framework for agency key performance indicators.
  - Increases the availability and visibility of SAMHSA information.
  - Offers tools for analysis, visualization and mash ups with other sources.
Case Study: Multiple categorization schemes (3)

SAMHSA Store Taxonomy facets
Agenda

- Interoperability
- Named entities exercise
- Problems with metadata
- Two types of vocabularies
  - Dates, roles and topics
  - Types of semantic schemes
  - Tagging assets
  - Tools for automating tagging
There are two types of vocabularies

- Concept schemes – metadata schemes like Dublin Core
- Semantic schemes – value vocabularies like taxonomies, thesauri, ontologies, etc.
What is metadata?

- Metadata provides enough information for any user, tool, or program to find and use any piece of content.

<table>
<thead>
<tr>
<th>Complexity</th>
<th>Enabled Functionality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asset metadata – <em>Who</em>: Identifier, Creator, Title, Description, Publisher, Format, Contributor</td>
<td>Use metadata – <em>When &amp; How</em>: Date, Language, Rights</td>
</tr>
<tr>
<td>Subject metadata – <em>What, Where &amp; Why</em>: Subject, Type, Coverage</td>
<td>Relational metadata – <em>Links between and to</em>: Source, Relation</td>
</tr>
</tbody>
</table>

http://dublincore.org/documents/dces/
What is metadata

- Metadata provides enough information for any user, tool, or program to find and use any piece of content.

---

### Metadata Types

- **Asset metadata**
  - **Who:** Identifier, Creator, Title, Description, Publisher, Format, Contributor

- **Subject metadata**
  - **What, Where & Why:** Subject, Type, Coverage

- **Relational metadata**
  - **Links between and to:** Source, Relation

- **Use metadata**
  - **When & How:** Date, Language, Rights

---

**Enabled Functionality**

- Better navigation & discovery
- More efficient editorial process

---

But Dublin Core is a little more complicated

<table>
<thead>
<tr>
<th>Elements</th>
<th>Refinements</th>
<th>Encodings</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Identifier</td>
<td>Abstract</td>
<td>Box</td>
<td>Collection</td>
</tr>
<tr>
<td>2. Title</td>
<td>Access rights</td>
<td>DCMIType</td>
<td>Dataset</td>
</tr>
<tr>
<td>3. Creator</td>
<td>Alternative</td>
<td>DDC</td>
<td>Event</td>
</tr>
<tr>
<td>4. Contributor</td>
<td>Audience</td>
<td>IMT</td>
<td>Image</td>
</tr>
<tr>
<td>5. Publisher</td>
<td>Available</td>
<td>ISO3166</td>
<td>Interactive</td>
</tr>
<tr>
<td>6. Subject</td>
<td>Bibliographic citation</td>
<td>ISO639-2</td>
<td>Resource</td>
</tr>
<tr>
<td>7. Description</td>
<td>Conforms to</td>
<td>LCC</td>
<td>Moving Image</td>
</tr>
<tr>
<td>8. Coverage</td>
<td>Created</td>
<td>LCSH</td>
<td>Physical Object</td>
</tr>
<tr>
<td>9. Format</td>
<td>Date accepted</td>
<td>MESH</td>
<td>Service</td>
</tr>
<tr>
<td>10. Type</td>
<td>Date copyrighted</td>
<td>Period</td>
<td>Software</td>
</tr>
<tr>
<td>11. Date</td>
<td>Date submitted</td>
<td>Point</td>
<td>Sound</td>
</tr>
<tr>
<td>12. Relation</td>
<td>Education level</td>
<td>RFC1766</td>
<td>Sound</td>
</tr>
<tr>
<td>13. Source</td>
<td>Extent</td>
<td>RFC3066</td>
<td>Still Image</td>
</tr>
<tr>
<td>14. Rights</td>
<td>Has format</td>
<td>TGN</td>
<td>Text</td>
</tr>
<tr>
<td>15. Language</td>
<td>Has part</td>
<td>UDC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Has version</td>
<td>MESH</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Is format of</td>
<td>URI</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Is part of</td>
<td>W3CTDF</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Is referenced by</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Is replaced by</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Is required by</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Issued</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Is version of</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>License</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mediator</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Modified</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Provenance</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>References</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Replaces</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Requires</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rights holder</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spatial</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Table of contents</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Temporal</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Valid</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Application profile**: Schema which consists of data elements drawn from one or more namespaces, combined together by implementers, and optimized for a particular local application.
Dublin Core is the top vocabulary in the linked data cloud

<table>
<thead>
<tr>
<th>Vocabulary prefix</th>
<th>Vocabulary link</th>
<th>Number of usages in data sets</th>
<th>Data sets that use the vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>dc</td>
<td><a href="http://purl.org/dc/elements/1.1/">http://purl.org/dc/elements/1.1/</a></td>
<td>92 (31.19 %)</td>
<td>Data sets that use dc</td>
</tr>
<tr>
<td>foaf</td>
<td><a href="http://xmlns.com/foaf/0.1/">http://xmlns.com/foaf/0.1/</a></td>
<td>81 (27.46 %)</td>
<td>Data sets that use foaf</td>
</tr>
<tr>
<td>skos</td>
<td><a href="http://www.w3.org/2004/02/skos/core#">http://www.w3.org/2004/02/skos/core#</a></td>
<td>58 (19.66 %)</td>
<td>Data sets that use skos</td>
</tr>
<tr>
<td>geo</td>
<td><a href="http://www.w3.org/2003/01/geo/wgs84_pos#">http://www.w3.org/2003/01/geo/wgs84_pos#</a></td>
<td>25 (8.47 %)</td>
<td>Data sets that use geo</td>
</tr>
<tr>
<td>xhtml</td>
<td><a href="http://www.w3.org/1999/xhtml/vocab#">http://www.w3.org/1999/xhtml/vocab#</a></td>
<td>19 (6.44 %)</td>
<td>Data sets that use xhtml</td>
</tr>
<tr>
<td>akt</td>
<td><a href="http://www.aktors.org/ontology/portal#">http://www.aktors.org/ontology/portal#</a></td>
<td>17 (5.76 %)</td>
<td>Data sets that use akt</td>
</tr>
<tr>
<td>bibo</td>
<td><a href="http://purl.org/ontology/bibo/">http://purl.org/ontology/bibo/</a></td>
<td>14 (4.75 %)</td>
<td>Data sets that use bibo</td>
</tr>
<tr>
<td>mo</td>
<td><a href="http://purl.org/ontology/mo/">http://purl.org/ontology/mo/</a></td>
<td>13 (4.41 %)</td>
<td>Data sets that use mo</td>
</tr>
<tr>
<td>vcard</td>
<td><a href="http://www.w3.org/2006/vcard/ns#">http://www.w3.org/2006/vcard/ns#</a></td>
<td>10 (3.39 %)</td>
<td>Data sets that use vcard</td>
</tr>
<tr>
<td>sioc</td>
<td><a href="http://rdfs.org/sioc/ns#">http://rdfs.org/sioc/ns#</a></td>
<td>10 (3.39 %)</td>
<td>Data sets that use sioc</td>
</tr>
<tr>
<td>cc</td>
<td><a href="http://creativecommons.org/ns#">http://creativecommons.org/ns#</a></td>
<td>8 (2.71 %)</td>
<td>Data sets that use cc</td>
</tr>
<tr>
<td>geonames</td>
<td><a href="http://www.geonames.org/ontology#">http://www.geonames.org/ontology#</a></td>
<td>6 (2.03 %)</td>
<td>Data sets that use geonames</td>
</tr>
<tr>
<td>frbr</td>
<td><a href="http://purl.org/vocab/frbr/core#">http://purl.org/vocab/frbr/core#</a></td>
<td>6 (2.03 %)</td>
<td>Data sets that use frbr</td>
</tr>
<tr>
<td>xsd</td>
<td><a href="http://www.w3.org/2001/XMLSchema#">http://www.w3.org/2001/XMLSchema#</a></td>
<td>6 (2.03 %)</td>
<td>Data sets that use xsd</td>
</tr>
<tr>
<td>time</td>
<td><a href="http://www.w3.org/2006/time#">http://www.w3.org/2006/time#</a></td>
<td>5 (1.69 %)</td>
<td>Data sets that use time</td>
</tr>
<tr>
<td>event</td>
<td><a href="http://purl.org/NET/c4dm/event.owl#">http://purl.org/NET/c4dm/event.owl#</a></td>
<td>5 (1.69 %)</td>
<td>Data sets that use event</td>
</tr>
<tr>
<td>dbpedia</td>
<td><a href="http://dbpedia.org/resource/">http://dbpedia.org/resource/</a></td>
<td>5 (1.69 %)</td>
<td>Data sets that use dbpedia</td>
</tr>
<tr>
<td>gr</td>
<td><a href="http://purl.org/goodrelations/v1#">http://purl.org/goodrelations/v1#</a></td>
<td>4 (1.36 %)</td>
<td>Data sets that use gr</td>
</tr>
</tbody>
</table>

http://www4.wiwiss.fu-berlin.de/lodcloud/state/#structure
MDM model that integrates taxonomy and metadata

- Taxonomies, Vocabularies, Ontologies
- Dublin Core
- Per-Source Data Types, Access Controls, etc.

Source: Todd Stephens, BellSouth
Why Dublin Core? According to Todd Stephens ...

- Dublin Core is a de-facto standard across many other systems and standards
  - RSS (1.0), OAI (Open Archives Initiative), SEMI E36, etc.
  - Inside organizations – ECMS, SharePoint, etc.
- Mapping to DC elements from most existing schemes is simple.
- Metadata already exists in enterprise applications
  - Windchill, OpenText, MarkLogic, SAP, Documentum, MS Office, SharePoint, Drupal, etc.
Agenda

- Interoperability
- Named entities exercise
- Problems with metadata
- Two types of vocabularies
- Dates, roles and topics
- Types of semantic schemes
- Tagging assets
- Tools for automating tagging
## Dates, roles and topics

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
<th>Set By</th>
</tr>
</thead>
<tbody>
<tr>
<td>date.added</td>
<td>Date the asset was first added to the DAM.</td>
<td>DAM</td>
</tr>
<tr>
<td>date.lastModified</td>
<td>Date the asset was last reviewed for accuracy and relevance. Used for provenance and to validate content or rights.</td>
<td>DAM</td>
</tr>
<tr>
<td>date.reviewed</td>
<td>Date the content was last reviewed for accuracy and relevance. Used for provenance, and to compute a future date to recheck the content.</td>
<td>DAM</td>
</tr>
<tr>
<td>date.nextReviewed</td>
<td>Date of next scheduled review for accuracy and relevance.</td>
<td>Rule</td>
</tr>
<tr>
<td>date.embargoed</td>
<td>Date and time that content is scheduled to become available on the site. Content can be prepared in advance and system will push it out once the embargo date is reached.</td>
<td>Manual</td>
</tr>
<tr>
<td>date.subject</td>
<td>Date of the event, data, or other information depicted in the asset. Used for search and recall purposes. (This is not the date the asset was uploaded or last updated).</td>
<td>Manual</td>
</tr>
</tbody>
</table>
Dublin Core dates

- “A date associated with an event in the life cycle of the resource”
- Woefully underspecified.
- Typically the publication or last modification date.
- Best practice: YYYY-MM-DD

### Refinements
- Created
- Valid
- Available
- Issued
- Modified
- Date Accepted
- Date Copyrighted
- Date Submitted

### Encodings
- DCMI Period
- W3C DTF (Profile of ISO 8601)
## Dates, roles and topics

<table>
<thead>
<tr>
<th>Role</th>
<th>Description</th>
<th>Admin</th>
<th>Add</th>
<th>Edit</th>
<th>Delete</th>
<th>Approve</th>
<th>Review</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administrator</td>
<td>Technical administration of the DAM. Generally allowed to do anything, to keep the system running and up-to-date.</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Approver</td>
<td>Senior DAM staff with the authority to approve assets for publication. In small shops Contributors may also be Approvers. Larger shops, and those using outsider contractors will have many Contributors but just a few Approvers.</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Contributor</td>
<td>Editorial staff with authority to contribute new assets to the DAM. Their work must be approved by an Approver before it can be published. Administrators have the authority to approve content for publication, but only as an exception not the rule.</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
</tbody>
</table>
**Dates, roles and topics**

**Descriptive Categories**
(Assessment)

- **Chronology**
  - Historic
  - Before 1940
  - 1940's
  - 1950's
  - 1960's
  - 1970's
  - 1980's
  - 1990's
  - 2000's
  - 2010's

- **Genre**
  - Holiday
  - Celebration
  - Occasion
  - Travel

- **People Type**
  - Family (General)
  - Friends (General)
  - Undetermined

- **Event Type**
  - ... (details not visible)

- **Key Event**
  - ... (details not visible)

- **Key Object**
  - ... (details not visible)

- **Key Location**
  - ... (details not visible)

- **Key Individual**
  - ... (details not visible)

- **Age**
  - Newborn
  - Baby
  - Toddler
  - Pre-school
  - Elementary School
  - Adolescent
  - Young Adult
  - Middle Age
  - Elderly

---

*Taxonomy Strategies* The business of organized information
Agenda

- Interoperability
- Named entities exercise
- Problems with metadata
- Two types of vocabularies
- Dates, roles and topics
- Types of semantic schemes
- Tagging assets
- Tools for automating tagging
Semantic Schemes: Simple to Complex

A set of words/phrases that can be used interchangeably for searching. E.g., Hypertension, High blood pressure.

A list of preferred and variant terms.

A system for identifying and naming things, and arranging them into a classification according to a set of rules.

An arrangement of knowledge usually enumerated, that does not follow taxonomy rules. E.g., Dewey Decimal Classification.

A faceted taxonomy but uses richer semantic relationships among terms and attributes and strict specification rules.

After: Amy Warner. *Metadata and Taxonomies for a More Flexible Information Architecture*
# Key semantic relationships

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept</td>
<td>A unit of thought, an idea, meaning, or category of objects or events. A Concept is independent of the terms used to label it.</td>
</tr>
<tr>
<td>Preferred Label</td>
<td>A preferred lexical label for the resource such as a term used in a digital asset management system.</td>
</tr>
<tr>
<td>Alternate Label</td>
<td>An alternative label for the resource such as a synonym or quasi-synonym.</td>
</tr>
<tr>
<td>Broader Concept</td>
<td>Hierarchical link between two Concepts where one Concept is more general than the other.</td>
</tr>
<tr>
<td>Narrower Concept</td>
<td>Hierarchical link between two Concepts where one Concept is more specific than the other.</td>
</tr>
<tr>
<td>Related Concept</td>
<td>Link between two Concepts where the two are inherently &quot;related&quot;, but that one is not in any way more general than the other.</td>
</tr>
</tbody>
</table>
Some semantic relationships for IBM

Subject | Predicate  | Object                                           |
---------|------------|--------------------------------------------------|
lc:n79142877 | skos:prefLabel | International Business Machines Corporation  |
lc:n79142877 | skos:altLabel   | IBM                                             |
lc:n79142877 | skos:altLabel   | I.B.M.                                          |
The business of organized information

<table>
<thead>
<tr>
<th>Subject</th>
<th>Predicate</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>lc:sh85052028</td>
<td>skos:prefLabel</td>
<td>Fringe parking</td>
</tr>
<tr>
<td>lc:sh85052028</td>
<td>skos:altLabel</td>
<td>Park and ride systems</td>
</tr>
<tr>
<td>lc:sh85052028</td>
<td>skos:altLabel</td>
<td>Park and ride</td>
</tr>
<tr>
<td>lc:sh85052028</td>
<td>skos:altLabel</td>
<td>Park &amp; ride</td>
</tr>
<tr>
<td>lc:sh85052028</td>
<td>skos:altLabel</td>
<td>Park-n-ride</td>
</tr>
<tr>
<td>trt:Brddf</td>
<td>skos:prefLabel</td>
<td>Fringe parking</td>
</tr>
<tr>
<td>trt:Brddf</td>
<td>skos:altLabel</td>
<td>Park and ride</td>
</tr>
</tbody>
</table>

TRT = Transportation Research Thesaurus
Negotiations With Iran Over Nuclear Program May Resume

By STEVEN LEE MYERS and RICK GLADSTONE
Published: February 17, 2012

WASHINGTON — The United States and the European Union signaled on Friday that negotiations with Iran over its nuclear program could soon resume for the first time even as a telecommunications network vital to the global banking industry prepared to expel Iranian banks.

While senior American and European officials stopped short of declaring a diplomatic breakthrough, Iran dropped previously unacceptable preconditions for talks in a letter this week from its senior nuclear negotiator, Saeed Jalili, who declared his country’s “readiness for dialogue” at “the earliest possibility.”

After weeks of official bluster, ominous threats of military action and a worrysome run-up to a UN nuclear agency inspection, the new overtures mark a major shift in the tone and possibly substance of international efforts to curb Iran’s accelerating nuclear program.
Microformats require metadata and taxonomy

Google’s new right rail
Agenda

- Interoperability
- Named entities exercise
- Problems with metadata
- Two types of vocabularies
- Dates, roles and topics
- Types of semantic schemes
- Tagging assets
- Tools for automating tagging
The Tagging Problem

- How are we going to populate metadata elements with complete and consistent values?
- What can we expect to get from automatic classifiers?
Cheap and Easy Metadata

- Some fields will be constant across a collection
  - e.g., format, color, photographer or location
- In the context of a single collection those kinds of elements may add little value, but they add tremendous value when many collections are brought together into one place, and they are cheap to create and validate.
## 4 Indexing rules: How to use the taxonomy to tag content

<table>
<thead>
<tr>
<th>Rule</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Use specific terms</strong></td>
<td>Apply the most specific terms when tagging content. Specific terms can always be generalized, but generic terms cannot be specialized.</td>
</tr>
<tr>
<td><strong>Use multiple terms</strong></td>
<td>Use as many terms as necessary to describe <em>What the content is about</em> &amp; <em>Why it is important.</em></td>
</tr>
<tr>
<td><strong>Use appropriate terms</strong></td>
<td>Only fill-in the facets &amp; values that make sense. Not all facets apply to all content.</td>
</tr>
<tr>
<td><strong>Consider how content will be used</strong></td>
<td>Anticipate <em>how the content will be searched for</em> in the future, &amp; <em>how to make it easy to find it.</em> Remember that search engines can only operate on explicit information.</td>
</tr>
</tbody>
</table>
Tagging considerations

- Who should tag assets? Producers or editors?
- Taxonomy is often highly granular to meet task and re-use needs, but with detailed taxonomy it’s difficult to get complete and consistent tags.
- The more tags there are (and the more values for each tag), the more hooks to the content, but the more difficult it is to get completeness and consistency.
- If there are too many tags or tags are too detailed, producers will resist and use “general” tags (if available)
- Vocabulary is often dependent on originating department, but the lingo may not be readily understood by people outside the department (who are often the users).
Tagging considerations (2)

- Automatic classification tools exist, and are valuable, but results are not as good as people can do.
  - “Semi-automated” is best.
  - Degree of human involvement is a cost/benefit tradeoff.
Agenda

- Interoperability
- Named entities exercise
- Problems with metadata
- Two types of vocabularies
- Dates, roles and topics
- Types of semantic schemes
- Tagging assets

Tools for automating tagging
## Tools for tagging

<table>
<thead>
<tr>
<th>Vendor</th>
<th>Taxonomy Editing Tools</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autonomy</td>
<td>Autonomy Collaborative Classifier</td>
<td><a href="http://www.autonomy.com/content/Functionality/idol-functionality-categorization/index.en.html">www.autonomy.com/content/Functionality/idol-functionality-categorization/index.en.html</a></td>
</tr>
<tr>
<td>ConceptSearching</td>
<td>ConceptSearching</td>
<td><a href="http://www.conceptsearching.com">www.conceptsearching.com</a></td>
</tr>
<tr>
<td>Mondeca</td>
<td>Intelligent Topic Manager</td>
<td><a href="http://www.mondeca.com/Products/ITM">www.mondeca.com/Products/ITM</a></td>
</tr>
<tr>
<td>Temis</td>
<td>Temis Luxid® for Content Enrichment</td>
<td><a href="http://www.temis.com/?id=201&amp;selt=1">www.temis.com/?id=201&amp;selt=1</a></td>
</tr>
</tbody>
</table>
**Taxonomy tagging tools**

Microsoft Office Properties are ubiquitous but rarely used.

An immature area— No vendors are in upper-right quadrant! No ECM vendors in this list. Tagging is a “best of breed” application.

High functionality /high cost products ($50-100K)

Microsoft Office Properties are ubiquitous but rarely used.

An immature area— No vendors are in upper-right quadrant! No ECM vendors in this list. Tagging is a “best of breed” application.

High functionality /high cost products ($50-100K)
Taxonomy tools and business intelligence

- No taxonomy tool vendors have connectors, custom APIs or other direct integrations with leading business intelligence tools.

- SAS acquired Teragram in 2010.
  - Teragram is primarily an OEM business, not integrated with SAS business intelligence products. But this is likely to change.

- Business Objects acquired Inxight in 2007, which was acquired by SAP in 2008.
  - Inxight is not evident in SAP business intelligence products.
What did you get out of this workshop?
Thank You

QUESTIONS

Joseph Busch
jbusch@taxonomystrategies.com
(415) 377-7912
twitter.com/joebusch