Semantic Metadata’s Two Types of Vocabularies

Text Analytics World
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 is the semantic web?

- Making content web-accessible in a format that can be read and used by automated tools, so that people and machines can find, share and integrate information more easily.

- Some current examples, especially if they use semantics as the basis for that integration.
  - **Named Entity Services** – Named entity links to “authority files” are rendered by browsers.
  - **Dynamic Web Pages** – Content changes in response to different contexts or conditions.
  - **Personalization** – Tailoring to a user based on personal details or characteristics they provide (or are inferred based on other information).
  - **Mashups** – Combining data from more than one source into an integrated application.
NY Times linked data

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 assassination attempts...
Google’s new right rail – another form of linked data
Oracle events mashup with Google maps.
## Most widely used vocabularies in the linked data cloud
(as of 9/11/2011)

<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
Linked data cloud characteristics  (as of 9/11/2011)

http://www4.wiwiss.fu-berlin.de/lodcloud/state/#structure
Types of Vocabularies

- In the linked data cloud, there are two types of vocabularies:
  - Concept schemes – metadata schemes like Dublin Core
  - Semantic schemes – value vocabularies like taxonomies, thesauri, ontologies, etc.
Concept scheme: Dublin Core

- dc
  - identifier
  - title
  - contributor
  - publisher
  - subject
  - description
  - coverage
  - format
  - type
  - date
  - relation
  - source
  - rights
  - language
  - audience

- accessRights
  - isVersionOf
  - hasVersion
  - isRequiredBy
  - replaces
  - requires
  - isPartOf
  - hasPart
  - isReferencedBy
  - references
  - isFormatOf
  - hasFormat
  - conformsTo

- creator
  - valid
  - available
  - issued
  - modified
  - dateAccepted
  - dateCopyrighted
  - dateSubmitted

- spatial
  - temporal

- alternative
  - bibliographicCitation

- mediator
  - educationLevel

- extent
  - medium

- abstract
  - tableOfContents

Taxonomy Strategies The business of organized information
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.
MDM model that integrates taxonomy and metadata

Taxonomies, Vocabularies, Ontologies

Dublin Core

Enterprise Business Intelligence
- Portals
- Business Intelligence
- Search Engines
- Topic Maps

Semantic Metadata

Integration Metadata

Repository Collection

Structural Metadata

Asset Collection
- Systems
- Applications
- Interfaces
- Databases

Source: Todd Stephens, BellSouth

Per-Source Data Types, Access Controls, etc.
Another concept scheme: Schema.org

- aggregateRating
- brand
- color
- depth
- height
- involves
- isAccessoryOrSparePartFor
- isConsumableFor
- isRelatedTo
- isSimilarTo
- itemCondition
- logo
- manufacturer
- model
- offers
- productID
- releaseDate
- review
- sku
- weight
- width

- attendee
- duration
- endDate
- location
- offers
- performer
- startDate
- subEvent
- superEvent

- address
- aggregateRating
- containedIn
- event
- faxNumber
- geo
- globalLocationNumber
- interactionCount
- isicV4
- logo
- map
- openingHoursSpecification
- ion
- photo
- review
- telephone

- Creative Work
- Product
- Place
- Person
- Medical Entity
- Organization
- Event
- Intangible
## Value Vocabulary: Schema.org

<table>
<thead>
<tr>
<th>Administrative Area</th>
<th>Civic Structure</th>
<th>Landform</th>
<th>Landmarks</th>
<th>Local Business</th>
<th>Residence</th>
<th>Tourist Attraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>City</td>
<td>Airport</td>
<td>+ BodyOfWater</td>
<td></td>
<td>AnimalShelter</td>
<td>ApartmentComplex</td>
<td></td>
</tr>
<tr>
<td>Country</td>
<td>Aquarium</td>
<td>Continent</td>
<td>Mountain</td>
<td>+ AutomotiveBusiness</td>
<td>GatedResidence</td>
<td></td>
</tr>
<tr>
<td>State</td>
<td>Beach</td>
<td>Mountain</td>
<td>Volcano</td>
<td>ChildCare</td>
<td>Community</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BusStation</td>
<td></td>
<td></td>
<td>DryCleaningOrLaundry</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BusStop</td>
<td></td>
<td></td>
<td>+ EmergencyService</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Campground</td>
<td></td>
<td></td>
<td>EmploymentAgency</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cemetery</td>
<td></td>
<td></td>
<td>+ Entertainment</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Crematorium</td>
<td></td>
<td></td>
<td>Business</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EventVenue</td>
<td></td>
<td></td>
<td>+ FinancialService</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FireStation*</td>
<td></td>
<td></td>
<td>+ FoodEstablishment</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ GovernmentBuilding</td>
<td></td>
<td></td>
<td>+ GovernmentOffice</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hospital*</td>
<td></td>
<td></td>
<td>+ HealthAndBeauty</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MovieTheater*</td>
<td></td>
<td></td>
<td>Business</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Museum</td>
<td></td>
<td></td>
<td>+ HomeAndConstruction</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MusicVenue</td>
<td></td>
<td></td>
<td>Business</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Park</td>
<td></td>
<td></td>
<td>+ InternetCafe</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ParkingFacility</td>
<td></td>
<td></td>
<td>Library</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PerformingArtsTheater</td>
<td></td>
<td></td>
<td>+ LodgingBusiness</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ PlaceOfWorship</td>
<td></td>
<td></td>
<td>+ MedicalOrganization</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Playground</td>
<td></td>
<td></td>
<td>+ ProfessionalService</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PoliceStation*</td>
<td></td>
<td></td>
<td>RadioStation</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RVPark</td>
<td></td>
<td></td>
<td>RealEstateAgent</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>StadiumOrArena*</td>
<td></td>
<td></td>
<td>RecyclingCenter</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SubwayStation</td>
<td></td>
<td></td>
<td>SelfStorage</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TaxiStand</td>
<td></td>
<td></td>
<td>ShoppingCenter</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TrainStation</td>
<td></td>
<td></td>
<td>+ Store</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Zoo</td>
<td></td>
<td></td>
<td>TelevisionStation</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TouristInformation</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Center</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TravelAgency</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Semantic Schemes: Simple to Complex

- **Synonym Ring**: A set of words/phrases that can be used interchangeably for searching. E.g., Hypertension, High blood pressure.
- **Controlled Vocabulary**: A list of preferred and variant terms.
- **Taxonomy**: A system for identifying and naming things, and arranging them into a classification according to a set of rules.
- **Classification Scheme**: An arrangement of knowledge usually enumerated, that does not follow taxonomy rules. E.g., Dewey Decimal Classification.
- **Thesaurus**: A tool that controls synonyms and identifies the semantic relationships among terms.
- **Ontology**: A faceted taxonomy but uses richer semantic relationships among terms and attributes and strict specification rules.

**Relationships**

- **Equivalence**
- **Hierarchy**
- **Associative**

After: Amy Warner. *Metadata and Taxonomies for a More Flexible Information Architecture*
Q: How do you share a vocabulary across (and outside of) the enterprise?
A: With standards

- **ISO 2788:1986** Guidelines for the Establishment and Development of Monolingual Thesauri
- **ISO 5964:1985** Guidelines for the Establishment and Development of Multilingual Thesauri
- **ISO 25964** (combines 2788 and 5964) Thesauri and Interoperability with other Vocabularies
- **Zthes** specifications for thesaurus representation, access and navigation
- **W3C SKOS** Simple Knowledge Organization System
Why SKOS? According to Alistair Miles …

- **Ease of combination** with other standards
  - Vocabularies are used in great variety of contexts.
    - E.g., databases, faceted navigation, website browsing, linked open data, spellcheckers, etc.
  - Vocabularies are re-used in combination with other vocabularies.

- **Flexibility and extensibility** to cope with variations in structure and style
  - Variations between types of vocabularies
    - E.g., list vs. classification scheme
  - Variations within types of vocabularies
Why SKOS? (2)

❖ **Publish managed vocabularies** so they can readily be consumed by applications
  ▪ Identify the concepts
    - What are the named entities?
  ▪ Describe the relationships
    - Labels, definitions and other properties
  ▪ Publish the data
    - Convert data structure to standard format
    - Put files on an http server (or load statements into an RDF server)

❖ **Ease of integration** with external applications
  ▪ Use web services to use or link to a published concept, or to one or more entire vocabularies.
    - E.g., [Google maps API], [NY Times article search API], [Linked open data]

❖ **A W3C standard** like HTML, CSS, XML… and RDF, RDFS, and OWL
## Semantic relationships

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Definition</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

<table>
<thead>
<tr>
<th>Subject</th>
<th>Predicate</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>lc:n79142877</td>
<td>skos:prefLabel</td>
<td>International Business Machines Corporation</td>
</tr>
<tr>
<td>lc:n79142877</td>
<td>skos:altLabel</td>
<td>IBM</td>
</tr>
<tr>
<td>lc:n79142877</td>
<td>skos:altLabel</td>
<td>I.B.M.</td>
</tr>
</tbody>
</table>
My company sells IBM’s XIV product

- **XIV**
- **XIV Storage System**
- **International Business Machines**

A high-end disk storage server designed to provide high performance, scalability, and availability in disk storage.

Originally developed by Israeli company XIV, which was acquired by IBM in 2007.
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>Use specific terms</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>Use multiple terms</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>Use appropriate terms</td>
<td>Only fill-in the facets &amp; values that make sense. Not all facets apply to all content.</td>
</tr>
<tr>
<td>Consider how content will be used</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>
Paper or web-based forms widely used:

- Distributed resource origination metadata tagging
- Centralized clean-up and metadata entry.

Source: CEN/ISSS Workshop on Dublin Core.
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.
## 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>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></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;sel=1">www.temis.com/?id=201&amp;sel=1</a></td>
</tr>
</tbody>
</table>
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.
- Business Objects acquired Inxight in 2007, which was acquired by SAP in 2008.
  - Inxight is not evident in SAP business intelligence products.
QUESTIONS?

Joseph A Busch, Principal

jbusch@taxonomystrategies.com
twitter.com/joebusch
415-377-7912
Semantic Metadata’s Two Types of Vocabularies

- Semantic metadata is metadata that is expressed using a standard syntax that can be commonly processed by applications and tools. This talk will focus on: the two types of vocabularies involved with semantic metadata, the class-attribute vocabulary, and the value vocabulary. We will explain why the most common vocabularies in the linked data cloud are Dublin Core, FOAF and SKOS; and how canonical lists of named entities (people, organizations, places, events and things) are being used to power new semantic services on web search engines, news websites and online shopping.