Controlled Vocabularies and the Dublin Core

Ron Daniel, Jr.

Tutorial 3: Vocabularies

13 September 2005

Agenda

9:00	Introduction
	Overview of talk
	Introductions
	Definitions
9:15	Which elements should use vocabularies, and which should use text?
9:30	Factoring “Subject” into Facets
9:50	Sources for Vocabularies
10:00	Maintaining Vocabularies
10:20	Q&A
10:30	Adjourn

Tutorial Description

- Title: Controlled Vocabularies and the Dublin Core
  Instructor: Ron Daniel
  Place: Aula de grados (5.1.A01)
  Date & Time: 14/09/05, from 9:00 to 10:30.

- Contents:
  The Dublin Core defines a number of metadata elements, but what about the values for those elements? Should they be unrestricted text values or come from pre-defined vocabularies? The answer, of course, is "it depends". During this tutorial we will discuss how to determine the appropriate approach for an organization's situation. We will also cover how pre-defined vocabularies should be sourced, structured, and maintained.

- This talk is oriented to an organizational intranet & knowledge management focus, not an academic & library focus.

Overall Context

Vocabulary development and maintenance is the LEAST of three problems:

- ***The Vocabulary Problem***: How are we going to build and maintain the lists of pre-defined values that can go into some of the metadata elements?
- ***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? What kind of error detection and error correction procedures do we need?
- ***The ROI Problem***: How are we going to use content, metadata, and vocabularies in applications to obtain business benefits?
  - More sales? Lower support costs? Greater productivity?
  - How much content? How big an operating budget?

Need to know the answer to the ROI Problem before solving the Vocabulary Problem.

Who we are: Ron Daniel, Jr.

- Over 15 years in the business of metadata & automatic classification
  - Principal, Taxonomy Strategies LLC
  - 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.

Taxonomy Strategies: Recent & current clients

- **Governmental**
  - Chelan County Public Utilities District
  - Commodity Futures Trading Commission
  - Defense Intelligence Agency
  - ERIC
  - Federal Aviation Administration
  - Federal Reserve Bank of Atlanta
  - Forest Service
  - GSA Office of Citizen Services (www.firstgov.gov)
  - Head Start
  - IMF
  - Infocomm Development Authority of Singapore
  - NASA (nasataxonomy.jpl.nasa.gov)
  - Small Business Administration
  - Social Security Administration
  - USDA Economic Research Service
  - USDA e-Government Program
- **NGOs**
  - CEN
  - IDEAlliance
  - OCLC
- **Commercial**
  - Allstate Insurance
  - BHP Billiton
  - Blue Shield of California
  - Debevoise & Plimpton
  - Halliburton
  - Hewlett Packard
  - Motorola
  - PeopleSoft
  - Pricewaterhouse Coopers
  - Siderean Software
  - Sprint
  - Time Inc.
- **Commercial subcontracts**
  - Agency.com – Top financial services
  - Critical Mass – Fortune 50 food retailer
  - Critical Mass – Fortune 50 hardware retailer
  - Deloitte Consulting – Big credit card
  - Gistics/OTB – Direct selling giant

Who are you? Tell us:

- Organization type:
  - Academic, Governmental, Non-Profit, Corporate, Other
- Organization size:
  - <10, 11..100, 101..1k, 1k..10k, 10..100k, > 100k
- Your role:
  - Information architect, Metadata designer, librarian, middle manager, CIO, student, IT Technical staff, Other
- The things you want to get from this workshop
  - More understanding of the use of vocabularies in Dublin Core
  - Details on how to build first vocabuary
  - Other

What is a controlled vocabulary? (Loose definition)

Hierarchical classification of things into a tree structure

- Animalia — Kingdom
- Chordata — Phylum
- Mammalia — Class
- Carnivora — Order
- Canidae — Family
- Canis — Genus
- C. familiari — Species

Linnaeus …

- 44-Office Equipment and Accessories and Supplies — Segment
- .12-Office Supplies — Family
- .17-Writing Instruments — Class
- .05-Mechanical pencils
- .06-Wooden pencils
- .07-Colored pencils — Commodity

UNSPSC …

Definitions

Term	Definition
Metadata Element	A ‘field’ for storing information about one piece of content. Examples: Title, Creator, Subject, Date, …
Metadata Value	The ‘contents’ of one Metadata Element. Values may be text strings, or selections from a predefined vocabulary.
Metadata Schema	A defined set of metadata elements. The Dublin Core is one schema.
Free Text Value	An unconstrained text metadata value. Some text values are constrained to follow a format (e.g. YYYY-MM-DD).
Vocabulary	A list of predefined values for a metadata element.
Controlled Vocabulary	A vocabulary with a defined and enforced procedure for its update.

Types of vocabularies

Vocabulary Type	Cplxty.	Description	Relation Type
Term List	1	Simple list of terms with no internal structure or relations.	None
Synonym Rings	2	List of sets of terms to regard as equivalent. Widely supported in search software.	Equivalence
Authority Files	3	List of names for known entities – people, ns, books, etc.	Reference
Classification Schemes		l arrangement of concepts.	Loose Hierarchy
Thesauri	5	Hierarchical arrangement of concepts plus supporting information and additional, non- hical, relations.	"Is-a" Hierarchy plus Loose Relations
Ontologi		ement of concepts and relations based odel of underlying reality – e.g. organs, oms, diseases & treatments in medicine.	Model-based Typed Relations

Rarely distinguished in practice

Search engine 'thesauri' may be synonym rings.

Vocabulary control

- The degree of control over a vocabulary is (mostly) independent of its type.
  - **Uncontrolled** – Anybody can add anything at any time and no effort is made to keep things consistent. Multiple lists and variations will abound.
  - **Managed** – Software makes sure there is a list that is consistent (no duplicates, no orphan nodes) at any one time. Almost anybody can add anything, subject to consistency rules. (e.g. File System Hierarchy)
  - **Controlled** – A documented process is followed for the update of the vocabulary. Few people have authority to change the list. Software may help, but emphasis is on human processes and custodianship. (e.g. Employee list)
- Term lists, synonym lists, … can be controlled, managed, or uncontrolled. Can't think of any unmanaged ontologies.

Pop Quiz

- How much control is needed over a vocabulary?

How much vocabulary control is really needed?

- Controlled vocabularies are frequently mentioned
  - That does not mean they are always necessary
  - Control comes at a cost, but can provide significant data quality benefits by reducing variations.
- Is this a well-controlled vocabulary?
  - No! It is an uncontrolled, but well-managed, term list
- Is this part of an appropriate solution to the ROI problem?
  - Yes! There is no budget to do ongoing control and QA

Source: http://del.icio.us/tag/

- Would this be appropriate for tracking royalty payments?
  - Of course not!

Agenda

9:00	Introduction
9:15	Which elements should use vocabularies, and which should use text?
9:30	Factoring “Subject” into Facets
9:50	Sources for Vocabularies
10:00	Maintaining Vocabularies
10:20	Q&A
10:30	Adjourn

Likelihood of Using Controlled Vocabularies for Dublin Core Elements

	(Virtually) Mandatory	Highly Likely	Maybe	Highly Unlikely	(Virtually) Impossible
Language	RFC 3066				
Format	IMT				
Coverage		ISO 3166			
Type		DCMI Type?			
Subject		Custom			
Creator			LDAP?		
Publisher			Custom		
Contributor			LDAP?		
Identifier			Custom		
Date				W3C DTF	
Rights					
Title					
Relation					
Source					
Description					

These four elements are the ones that take the most thought when defining a metadata schema

(Virtually) Mandatory: Format and Language

DC recommends specific best practices:

- Language: RFC 3066 (which works with ISO 639)
- Format: Internet Media Types (aka MIME)

These vocabularies are widely used throughout the Internet. If you want to do something else, it should be justified.

- Describing physical objects?
  - Use Extent and Medium refinements instead of Format.
- Regional (vs. National) dialects?
  - a) Why?
  - b) Consider a custom element in addition to standard Language

(Virtually) Impossible: Description

- Abstracts are not like subject codes

Highly Likely: Coverage, Type

DC recommends specific best practices:

- Coverage: ISO 3166
  - ISO 3166 should be used unless you have good reasons to use something else
  - Consider Getty Thesaurus of Geographic Names if you need cities, rivers, etc.
  - DC provides Encodings for both
- Type: DCMITypes
  - We do not think the DCMIType list is a best practice
  - No widely accepted type list exists, so a custom list is likely

Unlikely: Date, Title, Relation, Source

- Date: *Could* use a predefined list, but best practice is to regard it as a text field that conforms to the W3C Date & Time Format (W3C DTF).
- Rights: *Could* come from a predefined list of allowed usages, but unlikely. Typically just a copyright statement.
- Title: *Could* come from an authority list, but we have never seen that in a corporate context.
- Relation & Source: *Could* come from a list of known resources, but we have never seen that. Only useful in limited collections.

Maybe: Creator, Contributor, Publisher, Identifier

- Creator, Contributor *could* come from an "authority file"
  - LC NAF in library contexts
  - LDAP Directory in corporate contexts
    - Recommended where possible
    - Many exceptions where author is outside LDAP
  - "Contributor" is not recommended
- Publisher *could* come from an authority file
  - Org chart in corporate contexts – e.g. internal records management system.
  - May want to augment with partners, competitors, regulators, and interested third parties as part of a more general "Organization" field – e.g. a competitive intelligence portal
- Identifier *should* be a URI
  - Organization may manage these, but its typically a text field, not a controlled list.

Highly Likely: Subject (and extensions)

- Best practice: Use pre-defined subject schemes, not user-selected keywords.
  - DC Encodings (DDC, LCC, LCSH, MESH, UDC) most useful in library contexts.
  - Not useful for most corporate needs
- Recommended: Factor "Subject" into separate *facets*.
  - People, Places, Organizations, Events, Objects, Products & Services, Industry sectors, Content types, Audiences, Business Functions, Competencies, …
- Store the different facets in different fields
  - Use DC elements where appropriate (coverage, type, audience, …)
  - Extend with custom elements for other fields (industry, products, …)
  - "dc:Subject" is the element to hold what is left over after the main facets have been factored out

Agenda

9:00	Introduction
9:15	Which elements should use vocabularies, and which should use text?
9:30	Factoring “Subject” into Facets
	Non-DC Elements
9:50	Sources for Vocabularies
10:00	Maintaining Vocabularies
10:20	Q&A
10:30	Adjourn

DMOZ: A worst case example of a unified ‘subject’

- DMOZ has over 600k categories
- Most are a combination of common facets – Geography, Organization, Person, Document Type, …
- (e.g.) Top: Regional: Europe: Spain: Travel and Tourism: **Travel Guides**

Business	**Biotechnology & Pharmaceuticals**	**Education & Training**					
Regional	Europe	Ireland	Business & Economy	Employment	Health & Medical		
Reference	**Education**	**Colleges & Universities**	**North America**	**United States**	**Maryland**	**Columbia Union College**	**Athletics**
Reference	**Education**	**K-12**	**Home Schooling**	**Unschooling**	**Chats and Forums**		
Science	**Math**	**Academic Departments**	**South America**	**Colombia**			
Society	People	Women	Science & Technology	Mathematics			
Science	Social Sciences	Linguistics	Translation	Associations			
Business	Small Business	Finance	Accounting				
Business	Accounting	Firms	Directories				
Business	Employment	By Industry					
Business	Healthcare	Employment	Regional				

	Competency (discipline)	11
	Geography	9
	Audience	9
	Topic	7
	Organization	5
	Doc Type	4
	Industry	4
	Process	4

Why do we advocate a faceted approach?

- Power
  - 4 independent categories of 10 nodes = 10,000 nodes (10^4)
- Faster construction
  - Use existing taxonomies in specific fields
- Reduced maintenance cost
- More opportunity for data reuse
- Can be easier to navigate with appropriate UI

Agency	Form Type	Industry Impact	Jurisdiction
0001 Legislative	Application	**00** Generic	**Federal**
1000 Judicial	Approval	**11** Agriculture	**State**
1100 Executive	Claim	**21** Mining	**Local**
Office of Pres	Information request	**22** Utilities	**Other**
0003 Exec Depts	Information submission	**23** Construct	
1200 Agriculture	Instructions	**31-33** Manuf	
1300 Commerce	Legal filing	**42** Wholesale	
9700 Defense	Payment	**44-45** Retail	
9100 Education	Procurement	**48-49** Trans	
8900 Energy	Renewal	**51** Info	
7500 HHS	Reservation	**52** Finance	
7000 DHS	Service request	**54** Profession	
8600 HUD	Test	**55** Mgmt	
1400 Interior	Other input	**56** Support	
1500 Justice	Other transaction	**61** Education	
1600 Labor		**62** Health Care	
1900 State		**71** Arts	
6900 Transport		**72** Hospitality	
2000 Treasury		**81** Other Services	
3600 Veterans		**92** Public Admin	
Ind Agencies			
Intl Orgs			

60 nodes
24,000 combinations

US GSA eForms taxonomy

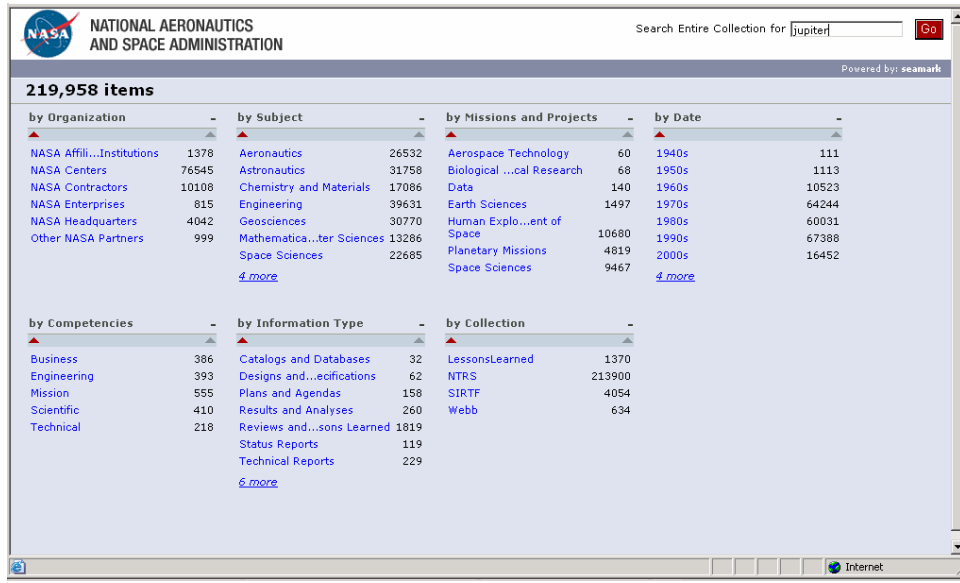
Agency	Form Type	Industry Impact	Jurisdiction	BRM Impact	Keyword Topic	Audience
0001 Legislative	Application	**00** Generic	**Federal**	**Citizen Srvcs**	Agriculture & food	**All**
1000 Judicial	Approval	**11** Agriculture	**State**	Social Srvs	Commerce	**General**
1100 Executive	Claim	**21** Mining	**Local**	Defense	Communica-tions	Citizen
Office of Pres	Information request	**22** Utilities	**Other**	Disasters	Education	Business
0003 Exec Depts	Information submission	**23** Construct		Econ Dev	Energy	Govt
1200 Agriculture	Instructions	**31-33** Manuf		Education	Env pro	Employee
1300 Commerce	Legal filing	**42** Wholesale		Energy	Foreign rels	Native
9700 Defense	Payment	**44-45** Retail		Env Mgmt	Govt	American
9100 Education	Procurement	**48-49** Trans		Law Enf	Health & safety	Nonresident
8900 Energy	Renewal	**51** Info		Judicial	Housing & comm dev	Tourist
7500 HHS	Reservation	**52** Finance		Correctional	Labor	**Special group**
7000 DHS	Service request	**54** Profession		Health	Law	
8600 HUD	Test	**55** Mgmt		Security	Named grps	
1400 Interior	Other input	**56** Support		Income Sec	National def	
1500 Justice	Other transaction	**61** Education		Intelligence	Nat resources	
1600 Labor		**62** Health Care		Intl Affairs	Recreation	
1900 State		**71** Arts		Nat Resour	Sci & tech	
6900 Transport		**72** Hospitality		Transport	Social pgms	
2000 Treasury		**81** Other Services		Workforce	Transport	
3600 Veterans		**92** Public Admin		Science		
Ind Agencies				**Delivery**		
Intl Orgs				**Support**		
				Management		

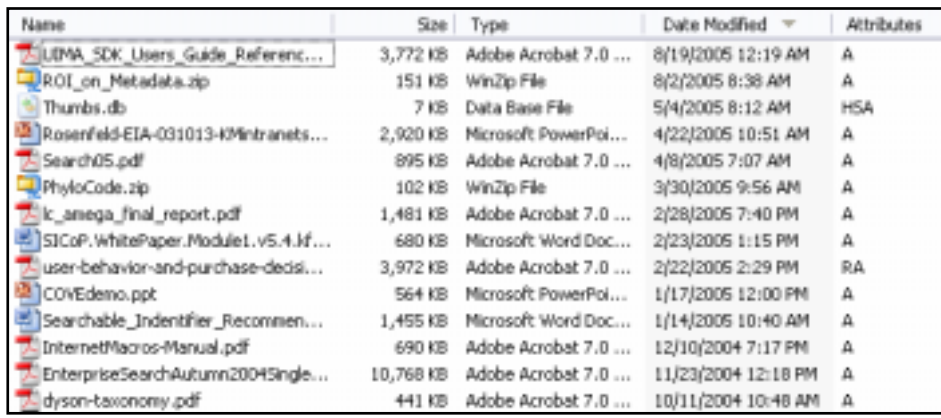
What do I do with all these facets?

- Either expose them directly in the user interface (post-coordinating)

  or

- Combine them in a minimal hierarchy (pre-coordination)
- Post-coordination takes software support, which may be fancy or basic.
- How many facets?
  - Log_{10}(#documents) as a guide





Define metadata specification, use DC elements for Integration

Element	Data Type	Length	Req. / Repeat	Source	Purpose
Asset Metadata					
Unique ID	I	**dc:identifier**	1	System supplied	Basic accountability
Recipe Title	S	**dc:title**	1	Licensed Content	Text search & results display
Recipe summary	S	**dc:description**	1	Licensed Content	Content
Main Ingredients	L	**X**	?	Main Ingredients vocabulary	Key index to retrieve & aggregate recipes, & generate shopping list
Subject Metadata					
Meal Types	L	**X**	*	Meal Types vocab	Browse or group recipes & filter search results
Cuisines	L	**X**	*	Cuisines	
Courses	L	**X**	*	Courses vocab	
Cooking Method	F	**X**	*	Cooking vocab	
Link Metadata					
Recipe Image	P	**dcterms:hasPart**	?	Product Group	Merchandize products
Use Metadata					
Rating	S		1	Licensed Content	Filter, rank, & evaluate recipes
Release Date	D	**dc:date**	1	Product Group	Publish & feature new recipes

dc:type="recipe", dc:format="text/html", dc:language="en"

Agenda

9:00	Introduction
9:15	Which elements should use vocabularies, and which should use text?
9:30	Factoring “Subject” into Facets
9:50	Sources for Vocabularies

- Which pre-existing ones should I use?
- How do I build my own? (Brief treatment)
- How do I decide on the necessary structure?

10:00	Maintaining Vocabularies
10:20	Q&A
10:30	Adjourn

Sources for 7 common vocabularies

cabulary	Definition	Potential Sources
Organization	Organizational structure.	FIPS 95-2, U.S. Government Manual, **Your organizational structure**, etc.
Content Type	Structured list of the various types of content being managed or used.	DC Types, AGLS Document Type, AAT Information Forms , Records management policy, etc.
Industry	Broad market categories such as lines of business, life events, or industry codes.	FIPS 66, SIC, **NAICS**, etc.
Location	Place of operations or constituencies.	FIPS 5-2, FIPS 55-3, **ISO 3166**, UN Statistics Div, US Postal Service, etc.
Function	Functions and processes performed to accomplish mission and goals.	**FEA Business Reference Model**, Enterprise Ontology, AAT Functions, etc.
Topic	Business topics relevant to your mission and goals.	**Federal Register Thesaurus**, NAL Agricultural Thesaurus, LCSH, etc.
Audience	Subset of constituents to whom a piece of content is directed or intended to be used.	GEM, ERIC Thesaurus, IEEE LOM, etc.
Products and Services	Names of products/programs & services.	ERP system, **Your products and services**, etc.

dc:publisher

dc:type

dc:coverage

dc:subject

dcterms:audience

Vocabulary construction

- The point of this talk is NOT to teach you how to build a vocabulary.
- The point of this talk is to help you structure things so that when you DO build vocabularies, you will succeed.
  - ROI and Tagging problems must be addressed so that you know the requirements
  - Metadata specification must exist, and fields should be factored so that the vocabularies are as concise as possible
  - Vocabularies should be taken from existing sources when possible, so that you have to do as little invention as possible
  - When you do build a new vocabulary, start small, relate it back to the content, and shift into a maintenance mode as soon as possible.

Five vocabulary construction rules

- URIs for node IDs, Publish vocabulary on web
- Modify metadata spec so vocabulary is indicated
- Gather data from multiple sources
  - Talk with users and experts
  - Analyze query logs and content
- Choose and arrange terms
  - Test and finalize first version
- Shift into maintenance mode

Seven practical rules for vocabularies

1. Incremental, extensible process that identifies and enables users, and engages stakeholders.
2. Quick implementation that provides measurable results as quickly as possible.
3. Not monolithic—has separately maintainable facets.
4. Re-uses existing IP as much as possible.
5. A means to an end, and not the end in itself .
6. Not perfect, but it does the job it is supposed to do—such as improving search and navigation.
7. Improved over time, and maintained.

Agenda

9:00	Introduction
9:15	Which elements should use vocabularies, and which should use text?
9:30	Factoring “Subject” into Facets
9:50	Sources for Vocabularies
10:00	Maintaining Vocabularies
10:20	Q&A
10:30	Adjourn

CV Business Processes

- Controlled Vocabularies must change, gradually, over time if they are to remain relevant
- Maintenance processes need to be specified so that the changes are based on rational cost/benefit decisions, with an awareness of their impact
- A team will need to maintain the vocabularies on a part-time basis
- Vocabulary team reports to some other steering committee

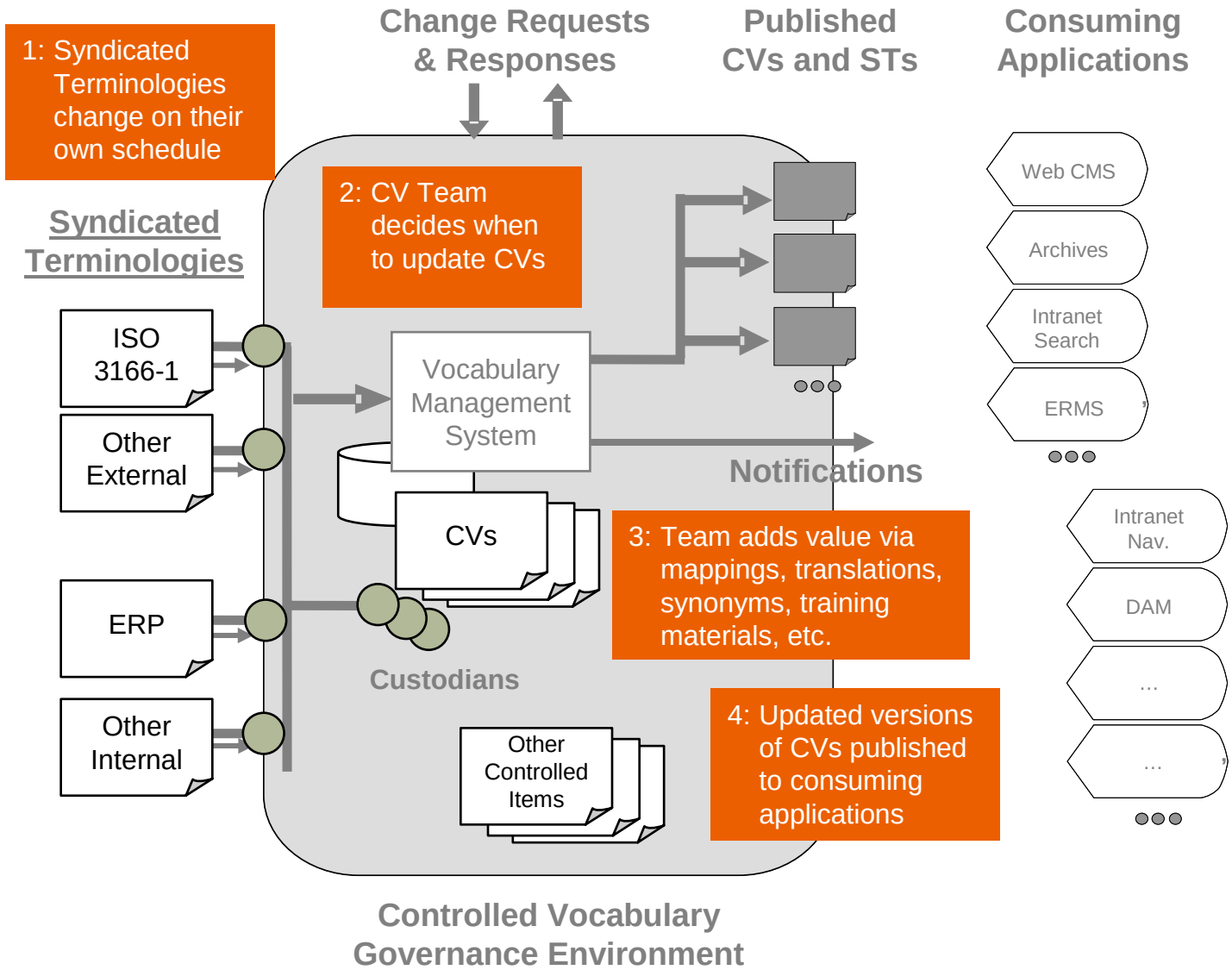
What could possibly go wrong with a little edit?

- ERP (Enterprise Resource Planning) team made a change to the product line data element in the product hierarchy.
- They did not know this data was used by downstream applications outside of ERP.
- An item data standards council discovered the error.
- If the error had not been identified and fixed, the company's sales force would not be correctly compensated.

"Lack of the enterprise data standards process in the item subject area has cost us at least 30 person days of just 'category' rework."

Source: Danette McGilvray, Granite Falls Consulting, Inc.

Definitions about the Controlled Vocabulary Governance Environment



Other Controlled Items

- CV Team will have additional items to manage:
  - Team Charter, Goals, Performance Measures
  - Editorial rules
  - Team processes
  - Tagger training materials (manual and automatic)
  - Outreach & ROI
    - Communication plan
    - Website
    - Presentations
    - Announcements
  - Vocabulary Roadmap

CV Governance: Generic team charter

- CV Team is responsible for maintaining:
  - The Taxonomy, a multi-faceted classification scheme
  - Associated materials, such as:
    - Editorial Style Guide
    - Training Materials
    - Metadata Standard
    - Team rules and procedures (subject to CIO review)
- Team evaluates costs and benefits of suggested change
- CV Team will:
  - Manage relationship between providers of source vocabularies and consumers of the CVs
  - Identify new opportunities for use of the CVs across the Enterprise to improve information management practices
  - Promote awareness and use of the CVs

Other Controlled Items - Editorial Rules

- To ensure consistent style, rules are needed
  - Akin to “Chicago Manual of Style”
- Issues commonly addressed in the rules:
  - Sources of Terms
  - Abbreviations
  - Ampersands
  - Capitalization
  - Continuations (More… or Other…)
  - Duplicate Terms
  - Fidelity
  - Hierarchy and Polyhierarchy
  - Languages and Character Sets
  - Length Limits
  - Numbers in Labels
  - “Other” – Allowed or Forbidden?
  - Plural vs. Singular Forms
  - Relation Types and Limits
  - … and many more
- Must also address issue of what to do when rules conflict – which are more important?

Rule Name	Editorial Rule
Use Existing Vocabularies	Other things being equal, reusing an existing vocabulary is preferred to creating a new one.
Ampersands	The character '&' is preferred to the word ‘and’ in Term Labels. Example: Use Type: “Manuals & Forms”, not “Manuals and Forms”.
Special Characters	Retain accented characters in Term Labels. Example: España
Serial comma	If a category name includes more than two items, separate the items by commas. The last item is separated by the character ‘&’ which IS NOT preceded by a comma. Example: “Education, Learning & Employment”, not “Education, Learning, & Employment”.
Capitalization	Use title case (where all words except articles are capitalized). Example: “Education, Learning & Employment” NOT “Education, learning & employment” NOT “EDUCATION, LEARNING & EMPLOYMENT”
…	…

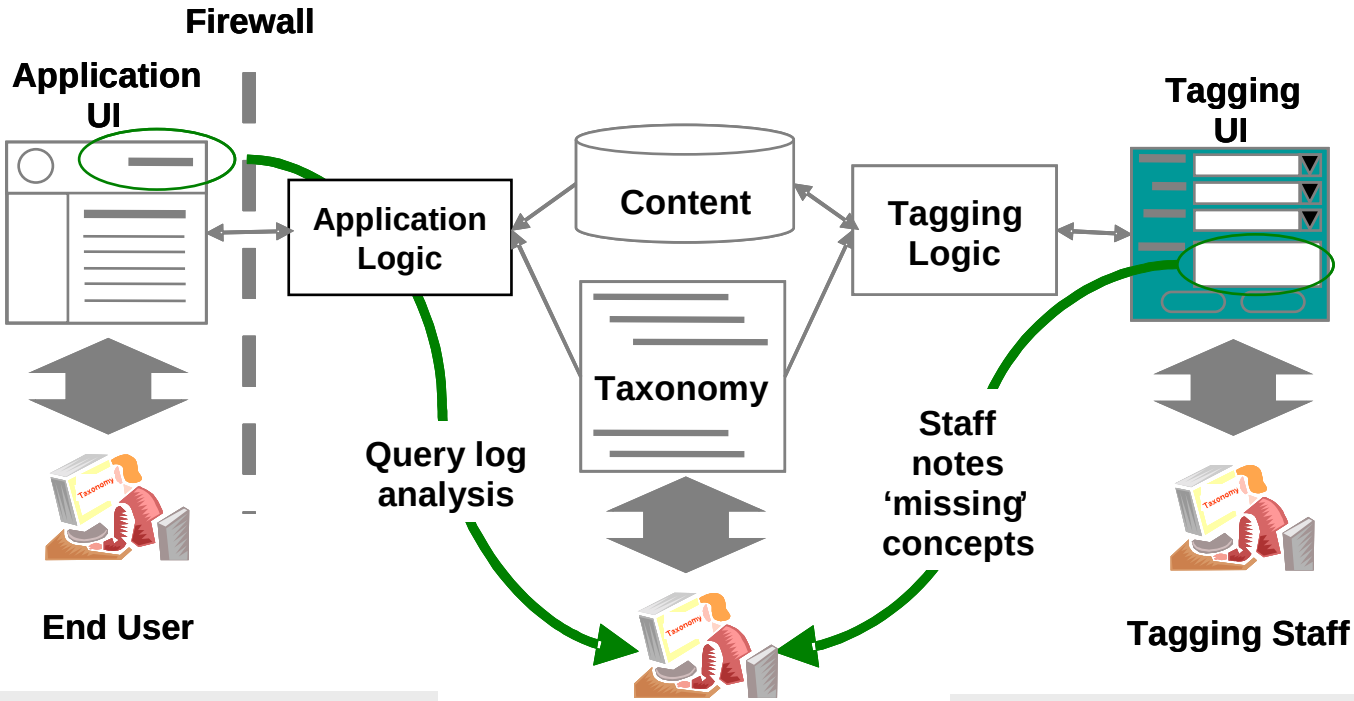
Roles in Two Taxonomy Governance Teams

- Executive Sponsor
  - Advocate for the taxonomy team
- Business Lead
  - Keeps team on track with larger business objectives
  - Balances cost/benefit issues to decide appropriate levels of effort
    - Specialists help in estimating costs
  - Obtains needed resources if those in team can't accomplish a particular task
- Technical Specialist
  - Estimates costs of proposed changes in terms of amount of data to be retagged, additional storage and processing burden, software changes, etc.
  - Helps obtain data from various systems
- Content Specialist
  - Team's liaison to content creators
  - Estimates costs of proposed changes in terms of editorial process changes, additional or reduced workload, etc.
  - **Small-scale Metadata QA Responsibility**
- Taxonomy Specialist
  - Suggests potential taxonomy changes based on analysis of query logs, indexer feedback
  - Makes edits to taxonomy, installs into system with aid of IT specialist
- Content Owner
  - Reality check on process change suggestions

Team structure at a different org.

- Business Lead
- Custodians
  - Responsible for content in a specific CV.
- Training Representative
  - Develops communications plan, training materials
- Work Practices Representative
  - Develops processes, monitors adherence
- IT Representative
  - Backups, admin of CV Tool
- Info. Mgmt. Representative
  - Provides CV expertise, tie-in with larger IM effort in the organization.

Taxonomy governance | Where changes come from



Recommendations by Editor

1. Small taxonomy changes (labels, synonyms)
2. Large taxonomy changes (retagging, application changes)
3. New "best bets" content

Team considerations

1. Business goals
2. Changes in user experience
3. Retagging cost

Agenda

9:00	Introduction
9:15	Which elements should use vocabularies, and which should use text?
9:30	Factoring “Subject” into Facets
9:50	Sources for Vocabularies
10:00	Maintaining Vocabularies
10:20	Q&A
10:30	Adjourn

Contact Info

Ron Daniel, Jr.

+1-925-368-8371

rdaniel@taxonomystrategies.com