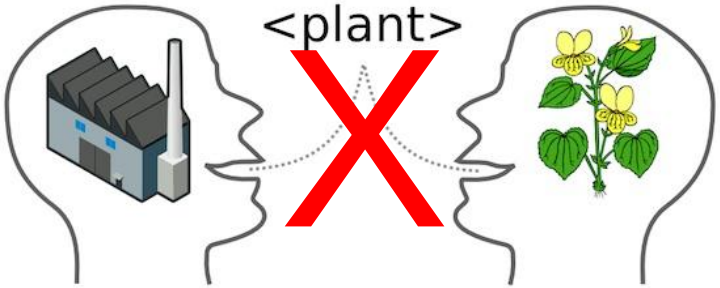
Metadata Interoperability



Agenda

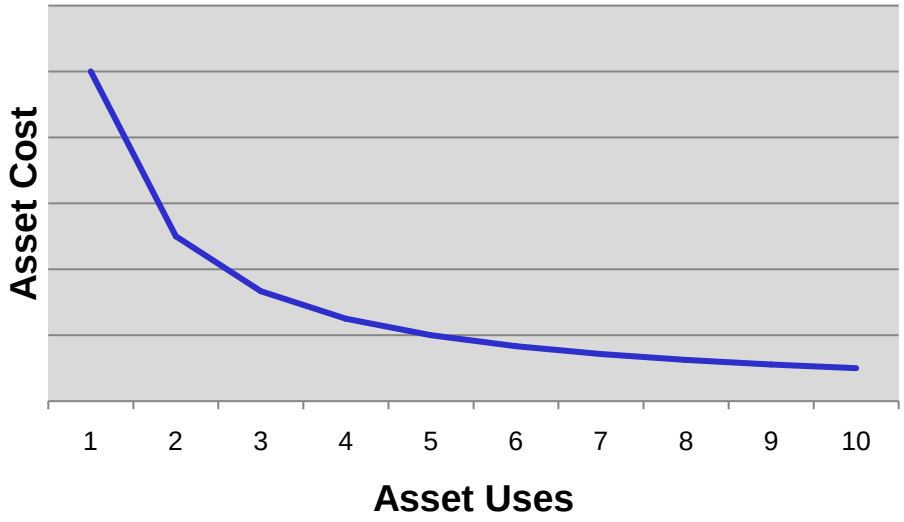
- Interoperability
- Dublin Core
- Dates, roles and topics
- Tagging assets
- Tools for automating tagging

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

- Content 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 content asset creation cost and increases the asset value.



Interoperability (2)

- If content assets are so important, why can't they be found?
  - They contain no searchable text, e.g., images.
  - They exist in different applications, file shares and/or desktops.
  - … Other reasons?
- When they are found why can't content 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)

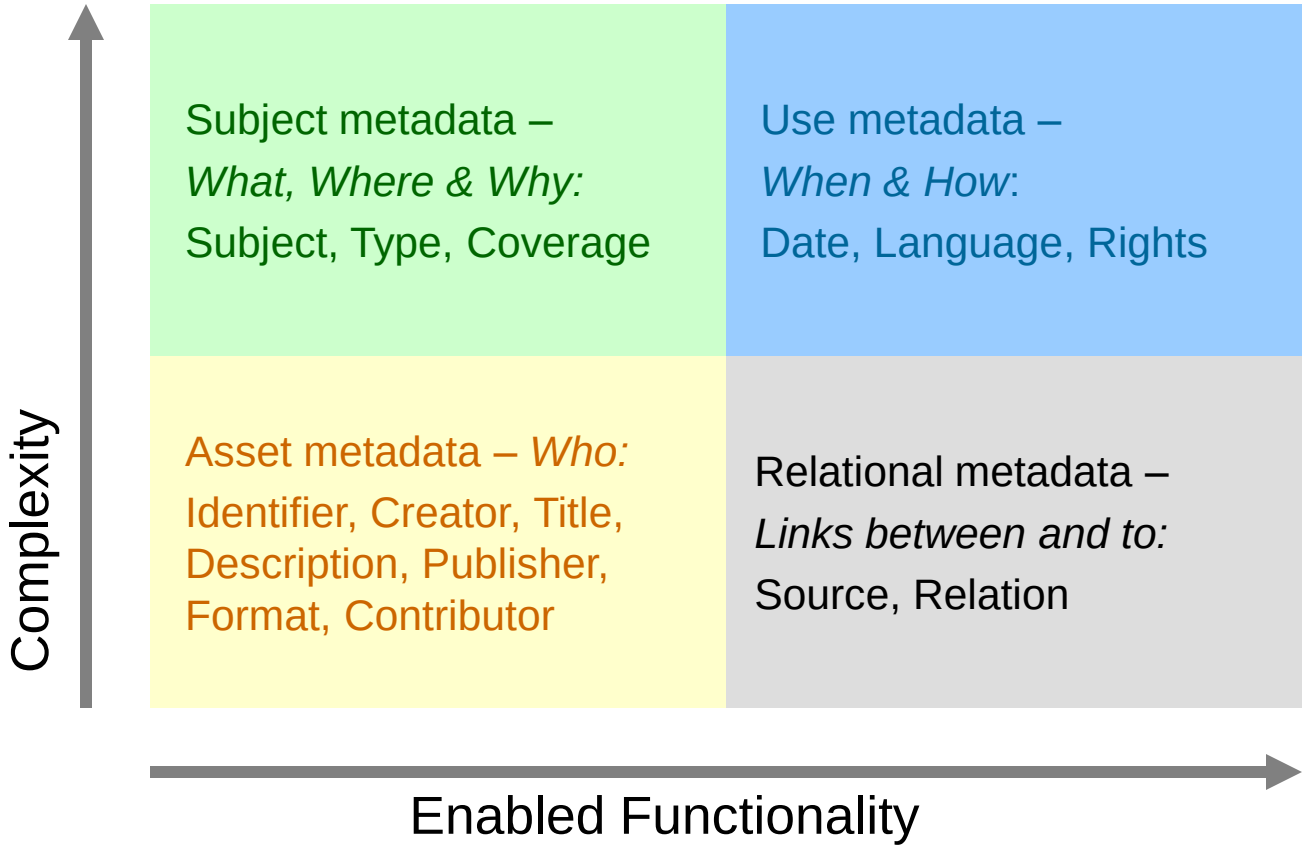
- Digital content assets are sourced from multiple applications and locations
  - Desktop applications
  - Enterprise content management applications
  - Product lifecycle management (PLM) applications
  - Product information management (PIM) applications
  - Third party contractors' systems
  - Other divisions and departments in your organization
  - Marketing and Communications servers
  - …Other applications and locations?

Interoperability vision

- I want to easily find any content 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., Author and Creator; Title and Caption; Location and Setting; etc.
  - How to align different vocabularies
    - E.g., CA and California; RiM and Research in Motion; etc.

What is metadata?

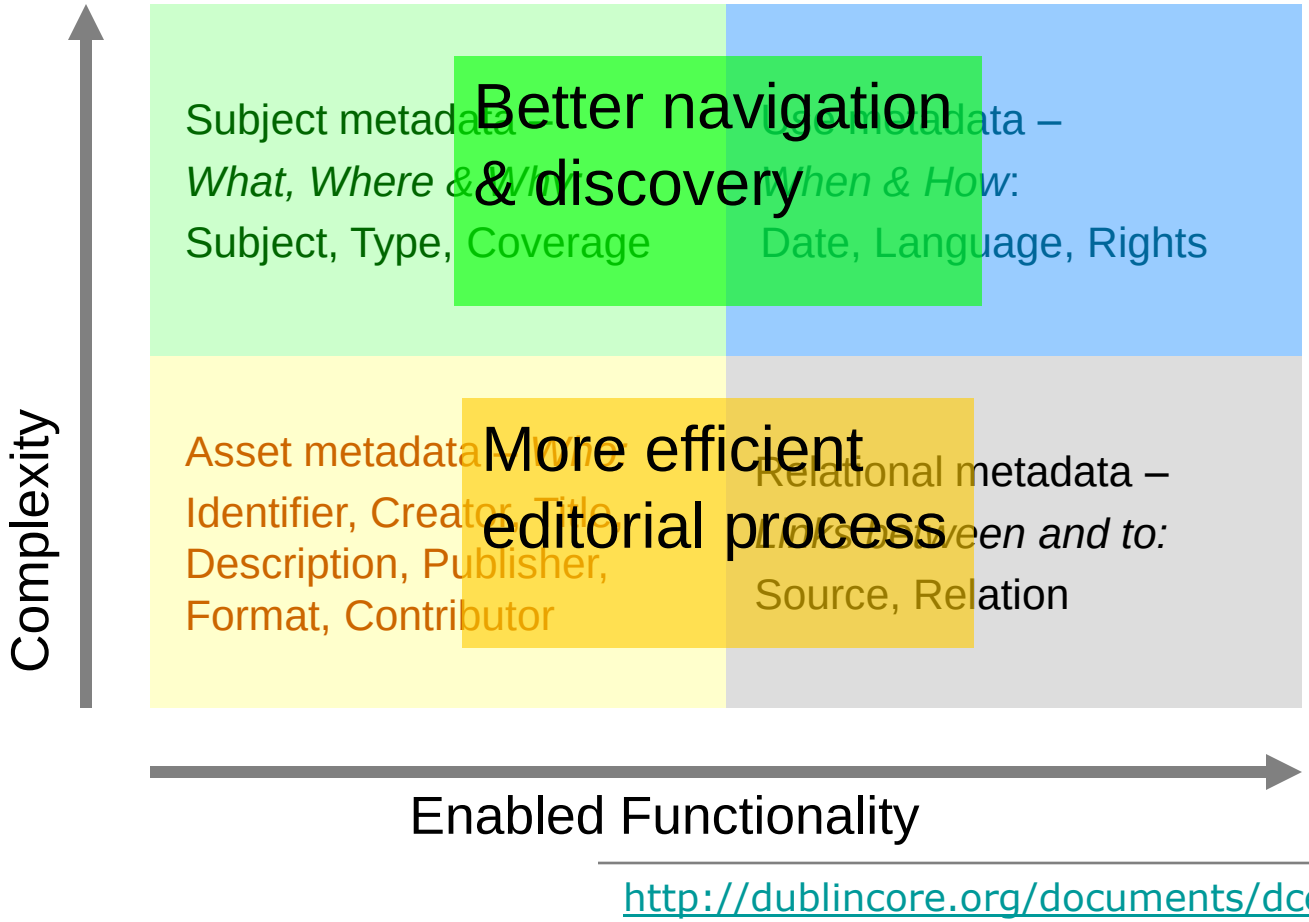
- Metadata provides enough information for any user, tool, or program to find and use any piece of content.



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.



http://dublincore.org/documents/dces/

But Dublin Core is a little more complicated

Elements

1. Identifier
2. Title
3. Creator
4. Contributor
5. Publisher
6. Subject
7. Description
8. Coverage
9. Format
10. Type
11. Date
12. Relation
13. Source
14. Rights
15. Language

Refinements

Abstract
Access rights
Alternative
Audience
Available
Bibliographic citation
Conforms to
Created
Date accepted
Date copyrighted
Date submitted
Education level
Extent
Has format
Has part
Has version
Is format of
Is part of
Is referenced by
Is replaced by
Is required by
Issued
Is version of
License
Mediator
Medium
Modified
Provenance
References
Replaces
Requires
Rights holder
Spatial
Table of contents
Temporal
Valid

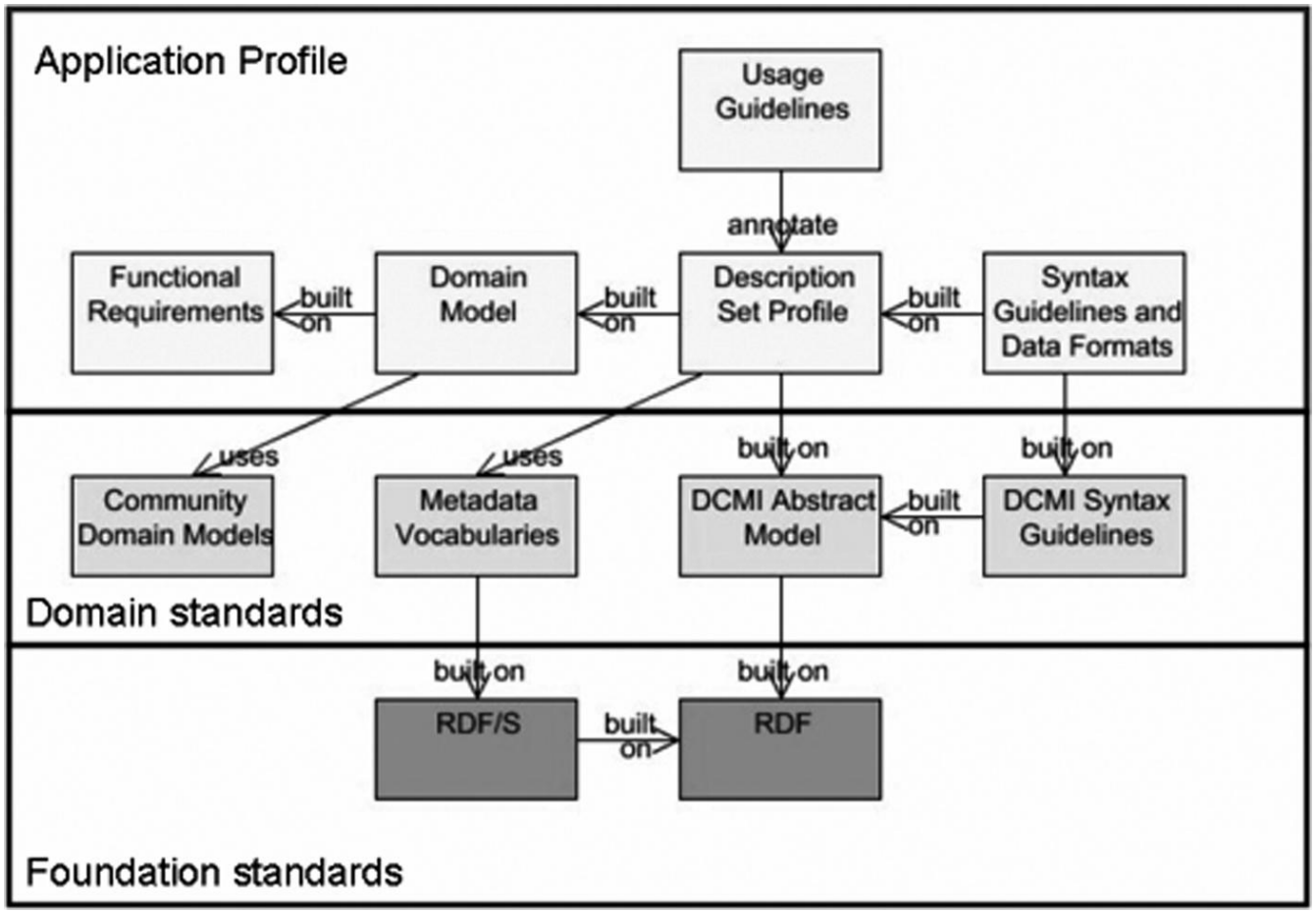
Encodings

Box
DCMIType
DDC
IMT
ISO3166
ISO639-2
LCC
LCSH
MESH
Period
Point
RFC1766
RFC3066
TGN
UDC
URI
W3CTDF

Types

Collection
Dataset
Event
Image
Interactive Resource
Moving Image
Physical Object
Service
Software
Sound
Still Image
Text

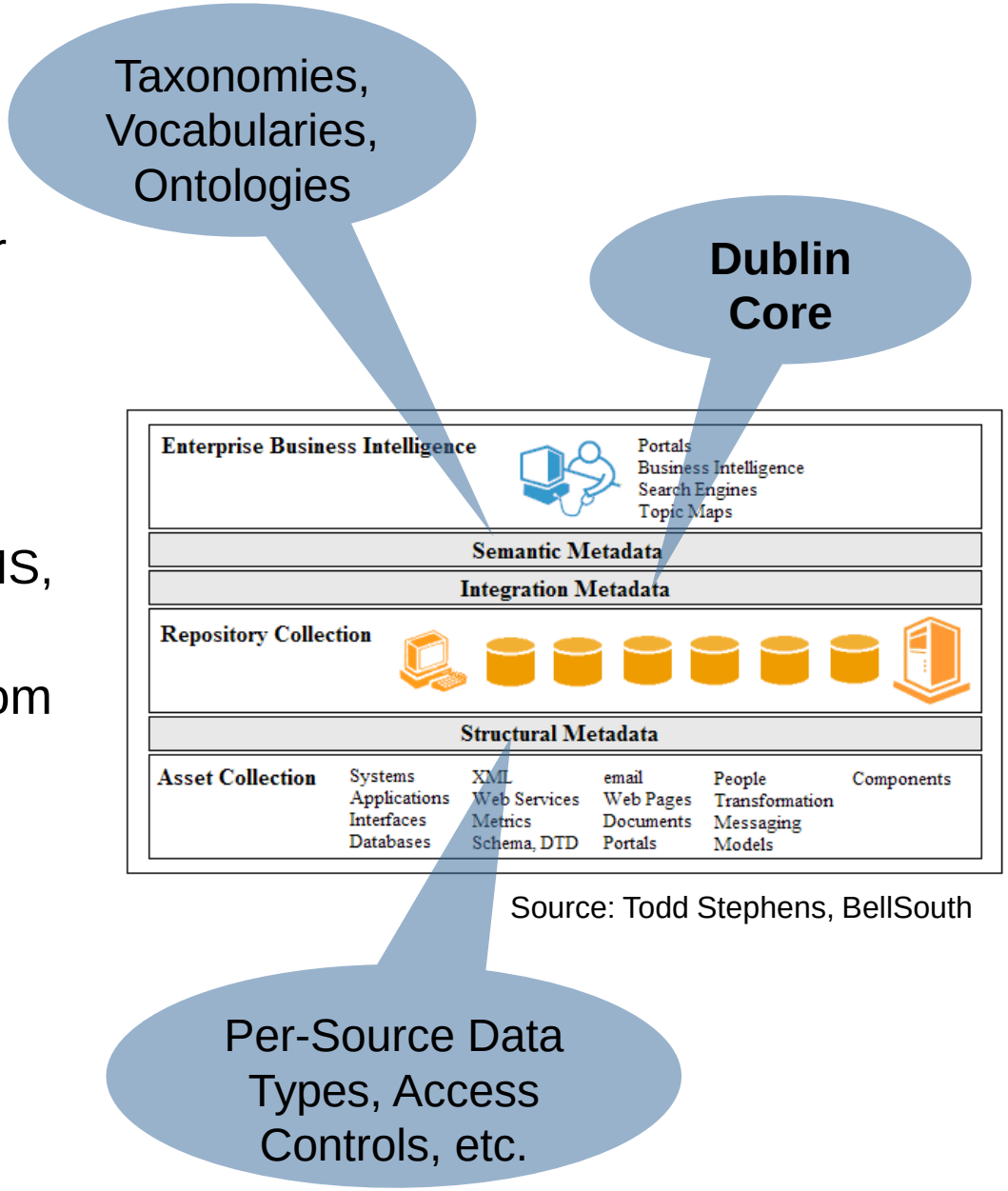
… and Dublin Core has gotten more abstract



DCAM (Dublin Core Abstract Model) Singapore Framework

So why Dublin Core?

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



Source: Todd Stephens, BellSouth

Dublin Core is the top vocabulary in the linked data cloud

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

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

Dates, roles and topics

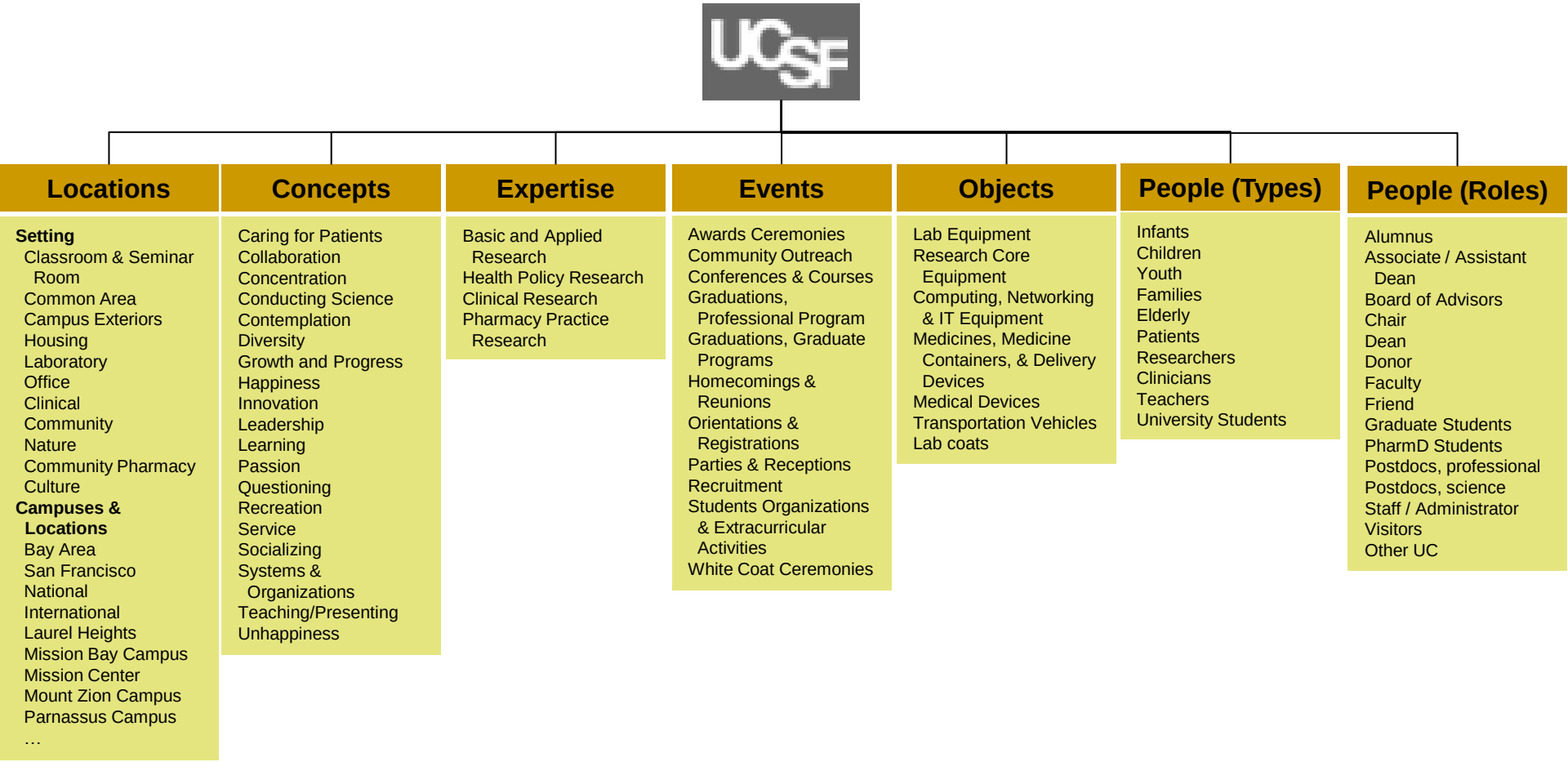
Explicit dates associated with an event in the life cycle of a resource.
Best practice: YYYY-MM-DD (*W3C DTF Profile of ISO 8601)*

Property	Description	Set By
date.added	Date the asset was first added to the DAM.	DAM
date.lastModified	Date the asset was last reviewed for accuracy and relevance. Used for provenance and to validate content or rights.	DAM
date.reviewed	Date the content was last reviewed for accuracy and relevance. Used for provenance, and to compute a future date to recheck the content.	DAM
date.nextReviewed	Date of next scheduled review for accuracy and relevance.	Rule
date.embargoed	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.	Manual
date.subject	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).	Manual

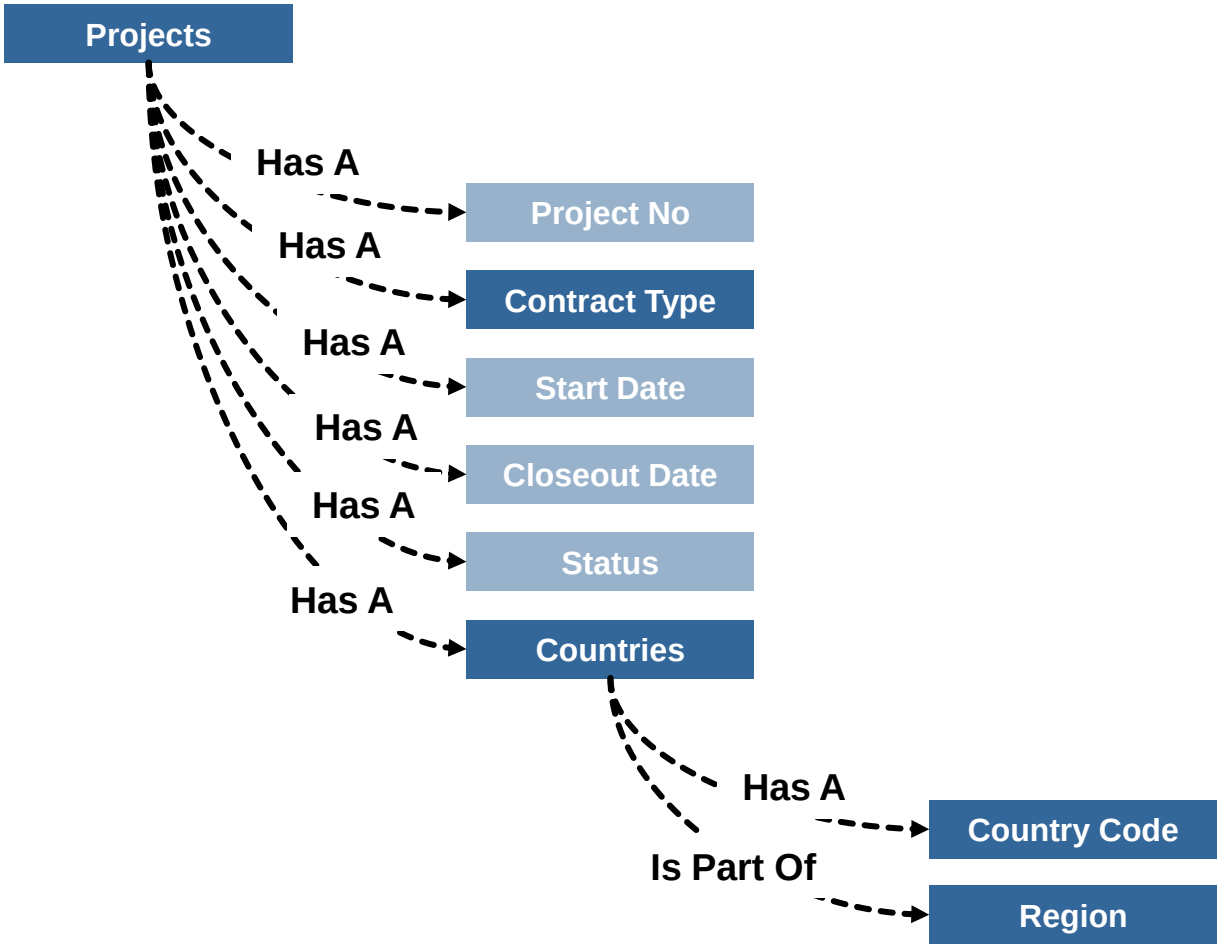
Dates, roles and topics

Role	Description	Admin	Add	Edit	Delete	Approve	Review
Administrator	Technical administration of the DAM. Generally allowed to do anything, to keep the system running and up-to-date.	Y	Y	Y	Y	Y	Y
Approver	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.	N	Y	Y	Y	Y	Y
Contributor	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.	N	Y	Y	N	N	Y

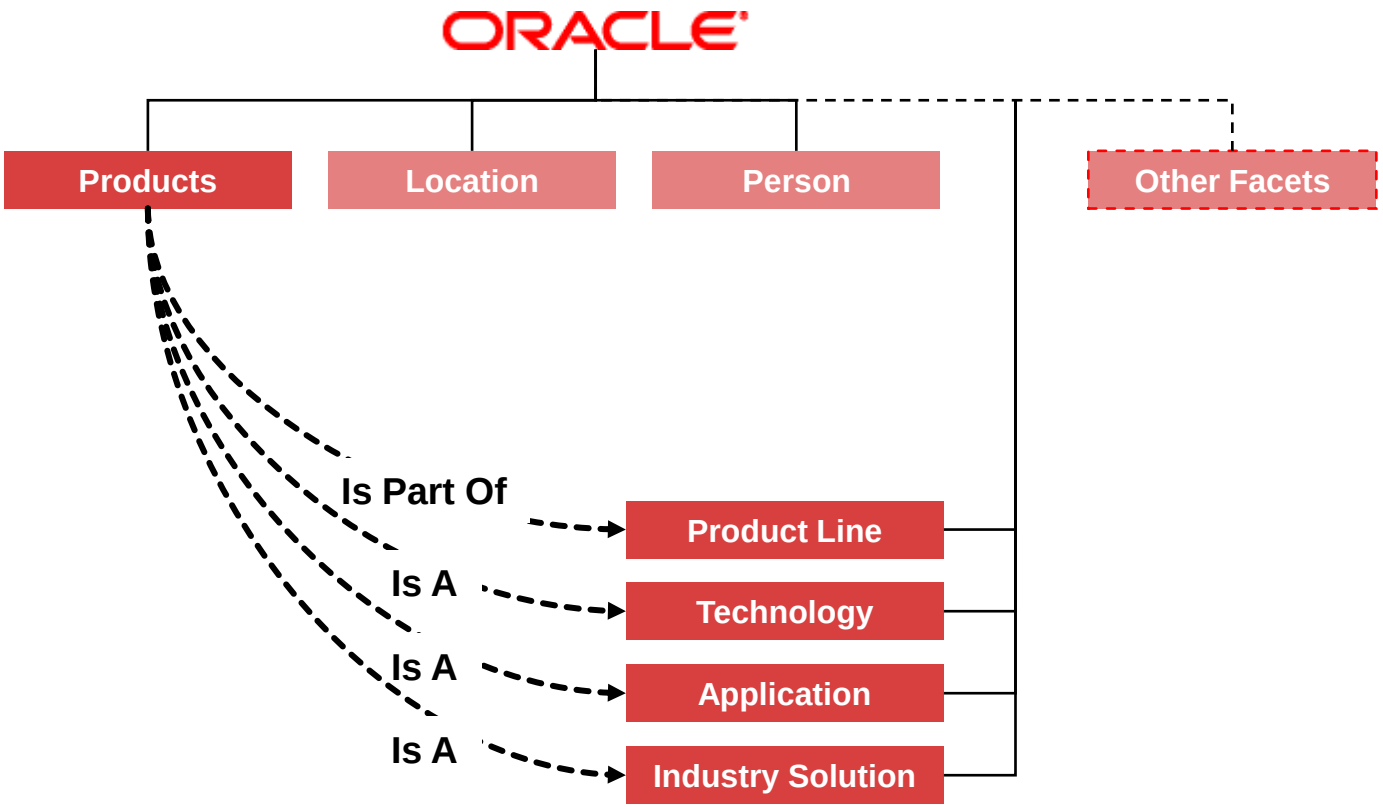
Dates, roles and topics



Semantic relationships among topics



Semantic relationships among topics



- Oracle events http://events.oracle.com/
- Oracle press releases http://pressroom.oracle.com/

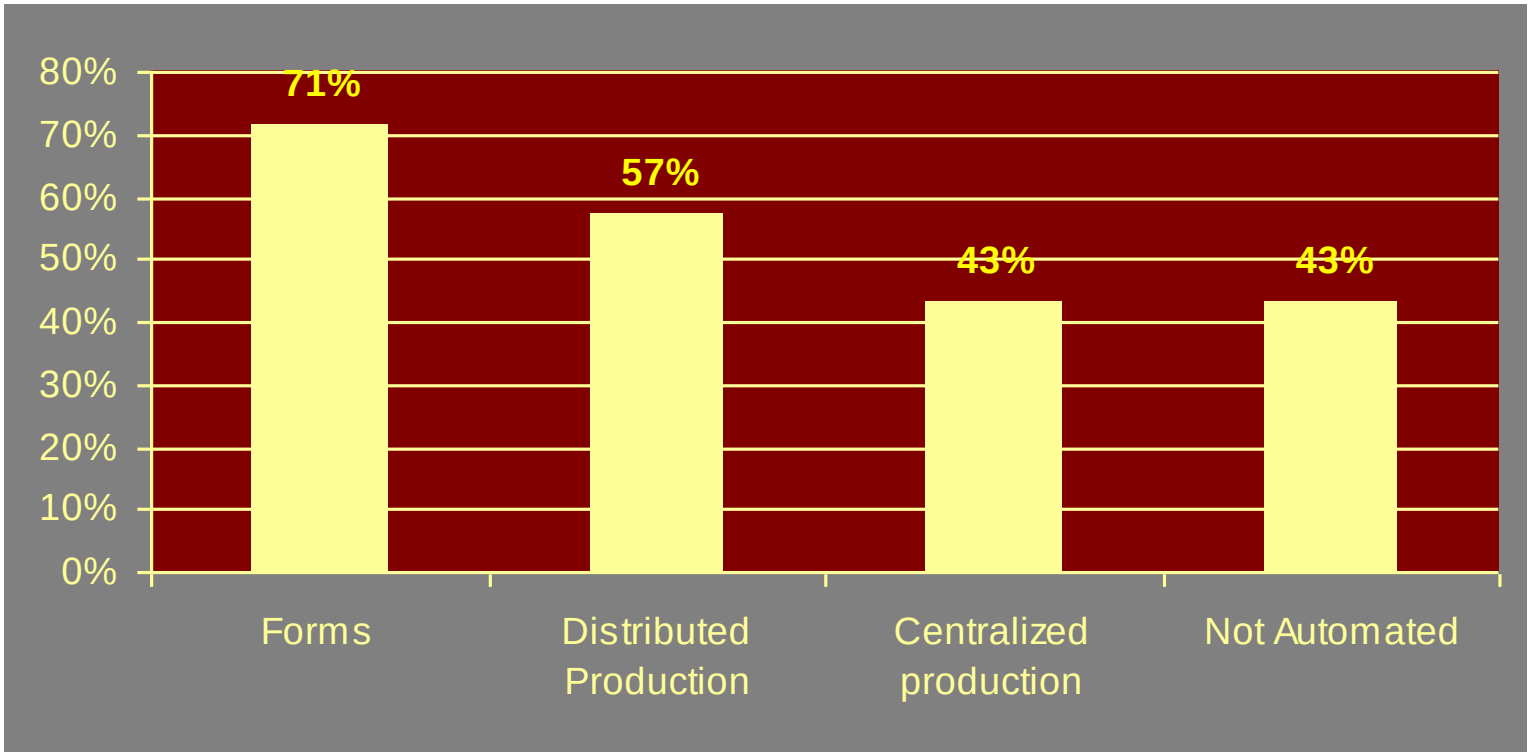
The Tagging Problem

- 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 group or collection of content items
  - E.g., project, project no, contract type, start date, closeout date, status, country and region
  - E.g., product, product line, technology, application and industry solution
- In the context of a single group or collection those kinds of elements may add little value, but they add tremendous value when many groups or collections are brought together into one place, and they are cheap to create and validate.

Methods used to create & maintain metadata



- Web-based forms are most 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).

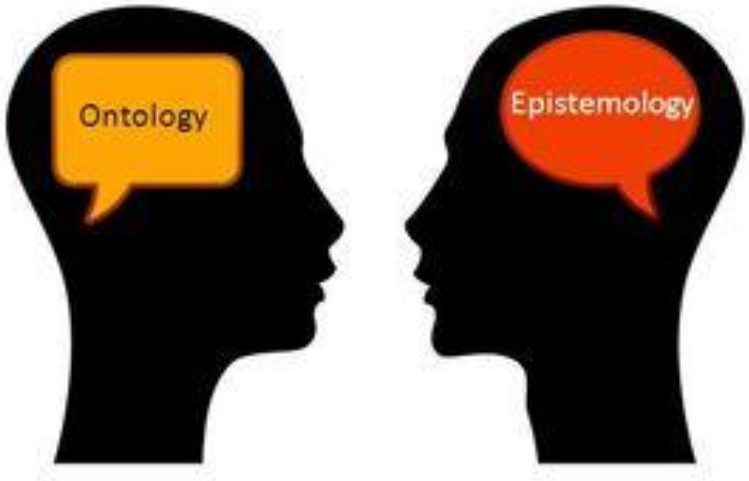
Tagging considerations (2)

- 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

Vendor	Taxonomy Editing Tools	URL
Autonomy an HP company	Autonomy Collaborative Classifier	www.autonomy.com/content/Functionality/idol-functionality-categorization/index.en.html
conceptSearching RETRIEVAL JUST GOT SMARTER	ConceptSearching	www.conceptsearching.com
Data Harmony Making Content Smarter	Data Harmony M.A.I.TM (Machine Aided Indexing)	www.dataharmony.com/products/mai.html
Office	Microsoft Office Properties	office.microsoft.com/en-us/access-help/view-or-change-the-properties-for-an-office-file-HA010354245.aspx?CTT=1
MONDECA	Intelligent Topic Manager	www.mondeca.com/Products/ITM
Nstein is now OPEN TEXT The Content Experts	nStein TME (Text Mining Engine)	www.nstein.com/en/products-and-technologies/text-mining-engine/
poolparty	PoolParty Extractor	poolparty.biz/products/poolparty-extractor/
Smartlogic	Semaphore Classification and Text Mining Server	www.smartlogic.com/home/products/semaphore-modules/classification-and-text-mining-server/overview
temis	Temis Luxid® for Content Enrichment	www.temis.com/?id=201&selt=1

24



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

Thank You

QUESTIONS