



Tutorial: Dublin Core, Metadata, and Growing the Semantic Web



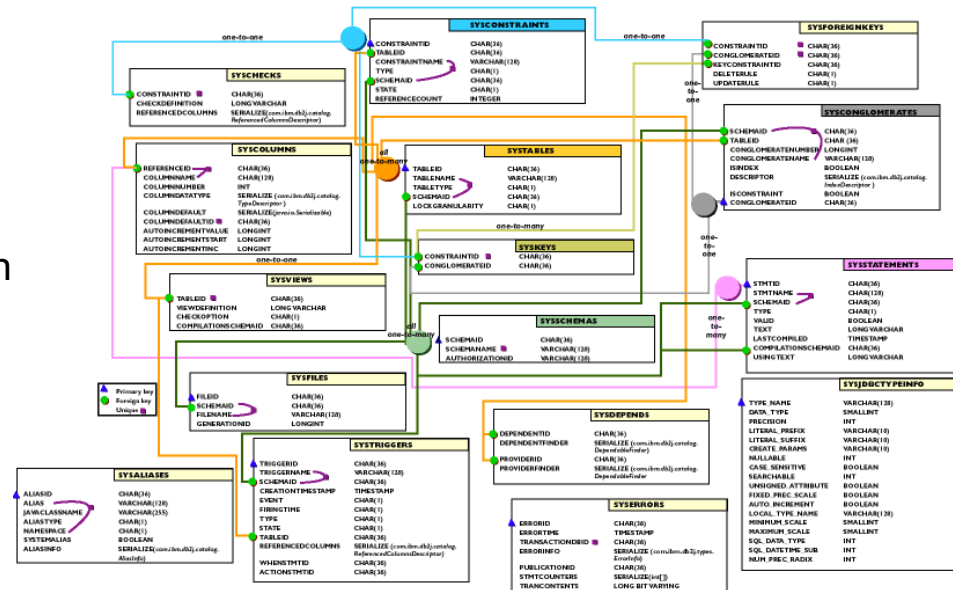
Metadata

- “Data about Data” – Amazingly vague
- Library & Information Science
 - Author/Title/Subject
 - Controlled Vocabularies for Subject Codes (e.g. Dewey)
 - Authority Files for Author Names
- Database
 - Tables/Columns/ Datatypes/Relationships
 - References for some values

530.1 H392b 1993	Black holes and baby universes and other essays Hawking, S.W. (Stephen W.) New York, N.Y. : Bantam Books, c. 1993 ix, 182 p. ; 24 cm
QC16.H33A3 1993	Library of Congress
530.1	93-8269
	AACR2 MARC

1. Hawking, SW 2. Cosmology
2. Science—Philosophy

- Other senses of the term
 - Statistics
 - Massive Storage
- Typologies of metadata
 - Administration/Preservation/Description
 - Asset/Use/Subject/Relation
 - Structural/Integration/Semantic



Metadata and Taxonomy Examples

Metadata

Field	Data Type / Source
Title	string
Creator	string
Identifier	URL
Date	date
Subject	category list

Taxonomy

The screenshot shows the DMOZ website interface. At the top, there is a green header with the DMOZ logo and the text "open directory project". Below the header, there are navigation links: "about dmoz", "suggest URL", "help", "link", and "editor login". A search bar is located in the center, with a "Search" button and a link to "advanced" search. Below the search bar, there is a grid of category links, each with a sub-link list:

- Arts**: [Movies](#), [Television](#), [Music](#)...
- Business**: [Jobs](#), [Real Estate](#), [Investing](#)...
- Computers**: [Internet](#), [Software](#), [Hardware](#)...
- Games**: [Video Games](#), [RPGs](#), [Gambling](#)...
- Health**: [Fitness](#), [Medicine](#), [Alternative](#)...
- Home**: [Family](#), [Consumers](#), [Cooking](#)...
- Kids and Teens**: [Arts](#), [School Time](#), [Teen Life](#)...
- News**: [Media](#), [Newspapers](#), [Weather](#)...
- Recreation**: [Travel](#), [Food](#), [Outdoors](#), [Humor](#)...
- Reference**: [Maps](#), [Education](#), [Libraries](#)...
- Regional**: [US](#), [Canada](#), [UK](#), [Europe](#)...
- Science**: [Biology](#), [Psychology](#), [Physics](#)...
- Shopping**: [Autos](#), [Clothing](#), [Gifts](#)...
- Society**: [People](#), [Religion](#), [Issues](#)...
- Sports**: [Baseball](#), [Soccer](#), [Basketball](#)...
- World**: [Deutsch](#), [Español](#), [Français](#), [Italiano](#), [Japanese](#), [Nederlands](#), [Polska](#), [Dansk](#), [Svenska](#)...

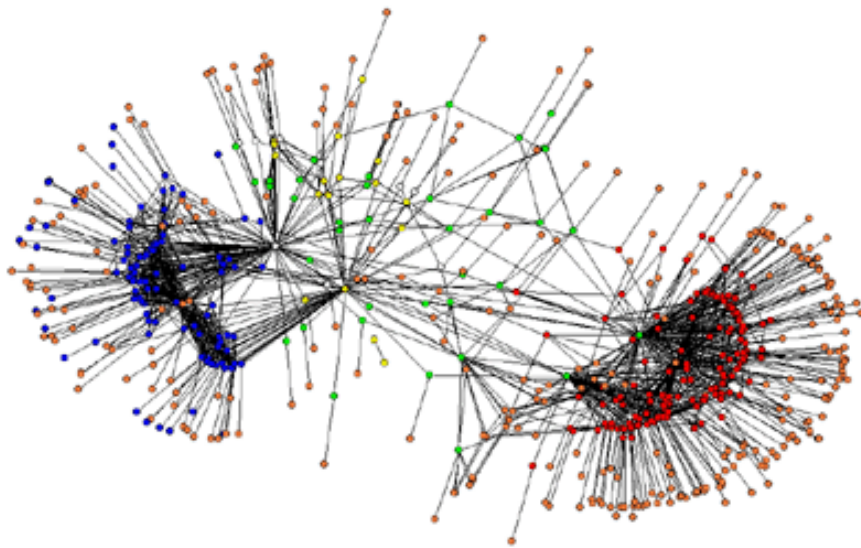
Dublin Core

Elements	Refinements	Encodings	Types
1. Identifier	Abstract	Is referenced by	Box
2. Title	Access rights	Is replaced by	DCMIType
3. Creator	Alternative	Is required by	DDC
4. Contributor	<i>Audience</i>	Issued	IMT
5. Publisher	Available	Is version of	ISO3166
6. Subject	Bibliographic citation	License	ISO639-2
7. Description	Conforms to	Mediator	LCC
8. Coverage	Created	Medium	LCSH
9. Format	Date accepted	Modified	MESH
10. Type	Date copyrighted	<i>Provenance</i>	Period
11. Date	Date submitted	References	Point
12. Relation	Education level	Replaces	RFC1766
13. Source	Extent	Requires	RFC3066
14. Rights	Has format	<i>Rights holder</i>	TGN
15. Language	Has part	Spatial	UDC
	Has version	Table of contents	URI
	Is format of	Temporal	W3CTDF
	Is part of	Valid	



Semantic Web

- The Semantic Web is built on top of the current Web, using RDF to assert machine-readable facts and inter-relations about Web resources.
 - The vision foresees a Network Effect from having many available datasets with machine-readable descriptions of their semantics.
- What is important about the Semantic Web is that new functionality comes from being able to take action on those facts and newly-created inter-relationships between formerly separate datasets.



<http://dmag.upf.es/livingsw/swgraph.htm>

Universitat Pompeu Fabra

Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://localhost:7056/course4/policy.rsp

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Sample query over the NASA Taxonomy

"Please find for me occasions in which policy issues posed threats to planetary missions. If possible, quote the appropriate passage from a mission report."

Note: the query is fixed, but the result is computed combining information from the NASA taxonomy and the HORM human and organizational risk research ontology.

Answers:

Mission Mars Climate Orbiter: **incident quote:** "Examining the current state of NASA's program and project management environment, the Board found that a significant infrastructure of processes and requirements already is in place to enable robust program and project management. However, these processes are not being adequately implemented within the context of "Faster, Better, Cheaper.""

Done

Courtesy Dean Allemang, Top Quadrant,
Robert Brummett, NASA HORM

Our assumptions on how the Semantic Web will evolve

1. We believe the Semantic Web will grow *slowly*, as the byproduct of integrating multiple datasets using their *existing metadata* to achieve goals that *justify the cost* of developing and testing an application.
2. Much of that current metadata is (or can be viewed as) *Dublin Core*.
3. Some *metadata hygiene* practices will make integration easier
 - Metadata hygiene will only be practiced if it offers benefits in the short term, or at least does not increase costs in the short term and the architects are aware of the long term.
4. Cyc-like description, mapping, and integration is a neat trick, but common fields and common controlled vocabularies are a lot more pragmatic.



Growth from Current Metadata

- Three types of existing metadata:
- Schema information (fields/columns, tables, relations, datatypes)
 - Quickest and easiest mapping between separate databases
 - Manual mapping not trivial, Cyc wants to (help) automate it.
 - Glosses over any oddities and outliers in the cells
 - Nice ROI, if you can deal with the bad QA
- Instance information (values in cells)
 - Adding new metadata to the instances in a large collection is a job beyond the time and patience of a Semantic Web researcher.
 - May be added through automated means, but QC follows the rule above.
 - Only gets created when someone has a budget to spend on it.
- Reference data (lists of values for certain cells)
 - Intermediate level of difficulty
 - Commonly need to manually map one list of values to another



Creator

- “An entity primarily responsible for making the content of the resource”
- In other words – Author, Photographer, Illustrator, ...
 - Potential refinements by creative role
 - Rarely justified
- Creators can be persons or organizations
- Key Point - Dealing with names is a big issue in data quality:
 - Ron Daniel
 - Ron Daniel, Jr.
 - Ron Daniel Jr.
 - R.E. Daniel
 - Ronald Daniel
 - Ronald Ellison Daniel, Jr.
 - Daniel, R.
- Name fields may contain other information
 - <dc:creator>Case, W. R. (NASA Goddard Space Flight Center, Greenbelt, MD, United States)</dc:creator>

Refinements

None

Encodings

None

Example – Name mismatches

- One of these things is not like the other:
 - Ron Daniel, Jr. and Carl Lagoze; “Distributed Active Relationships in the Warwick Framework”
 - Hojung Cha and Ron Daniel; “Simulated Behavior of Large Scale SCI Rings and Tori”
 - ✓ Ron Daniel; “High Performance Haptic and Teleoperative Interfaces”
- Differences may not matter
- If they do
 - This error cannot be reliably detected automatically
 - Authority files and an **error-correction** procedure are needed



Contributor

- “An entity responsible for making contributions to the content of the resource.”
- In practice – rarely used. Difficult to distinguish from Creator.

Refinements

None

Encodings

None



Publisher

- “An entity responsible for making the resource available”.
- Problems:
 - All the name-handling stuff of Creator.
 - Hierarchy of publishers (Bureau, Agency, Department, ...)

Refinements

None

Encodings

None



Title

- “A name given to the resource”.
- Issues:
 - Hierarchical Titles
e.g. Conceptual Structures:
Information Processing in
Mind and Machine (The
Systems Programming Series)
 - Untitled Works

Refinements

Alternative

Encodings

None



Date

- “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)

Identifier

- “An unambiguous reference to the resource within a given context”
- Best Practice: URL
- Future Best Practice: URI?
- Problems
 - Metaphysics
 - Personalized URLs
 - Multiple identifiers for same content
 - Non-standard resolution mechanisms for URIs

Refinements

Bibliographic Citation

Encodings

URI

Subject

- The topic of the content of the resource.
- Best practice: Use pre-defined subject schemes, not user-selected keywords.
- Factor “Subject” into separate *facets*.
 - People, places, organizations, events, objects, services
 - Industry sectors
 - Content types, audiences, functions
 - Topic
- Some of the facets are already defined in DC (Coverage, Type) or DCTERMS (Audience)

Refinements

None

Encodings

DDC

LCC

LCSH

MESH

UDC

Coverage

- “The extent or scope of the content of the resource”.
- In other words – places and times as topics.
- Key Point – Locations important in SOME environments, irrelevant in others.

Refinements

Spatial
Temporal

Encodings

Box (for Spatial)
ISO3166 (for Spatial)
Point (for Spatial)
TGN (for Spatial)
W3CTDF (for Temporal)

Description

- “An account of the content of the resource”.
- In other words – an abstract or summary
- Key Point – What’s the cost/benefit tradeoff for creating descriptions?
 - Quality of auto-generated descriptions is low
 - For search results, hit highlighting is probably better

Refinements

Abstract
Table of Contents

Encodings

None

Type

- “The nature or genre of the content of the resource”
- Best Current Practice: Create a custom list of content types, use that list for the values.
 - Try to avoid “image”, “audio”, and other format names in the list of content types, they can be derived from “Format”.
 - No broadly-acceptable list yet found.

Refinements

None

Encodings

DCMI Type

Format

- The physical or digital manifestation of the resource.
- In other words – the file format
- Best practice: Internet Media Types
- Outliers: File sizes, dimensions of physical objects

Refinements

Extent
Medium

Encodings

IMT



Language

- “A language of the intellectual content of the resource”.
- Best Practice: ISO 639, RFC 3066
- Dialect codes: Advanced practice

Refinements

None

Encodings

ISO639-2
RFC1766
RFC3066

Relation

- “A reference to a related resource”
- Very weak meaning – not even as strong as “See also”.
- Best practice: Use a refinement element and URLs.

Refinements

Is Version Of
Has Version
Is Replaced By
Replaces
Is Required By
Requires
Is Part Of
Has Part
Is Referenced By
References
Is Format Of
Has Format
Conforms To

Encodings

URI

Source

- “A reference to a resource from which the present resource is derived”
- Original intent was for derivative works
- Frequently abused to provide bibliographic information for items extracted from a larger work, such as articles from a Journal

Refinements

None

Encodings

URI



Rights

- “Information about rights held in and over the resource”
- Could be a copyright statement, or a list of groups with access rights, or ...

Refinements

*Access Rights
License*

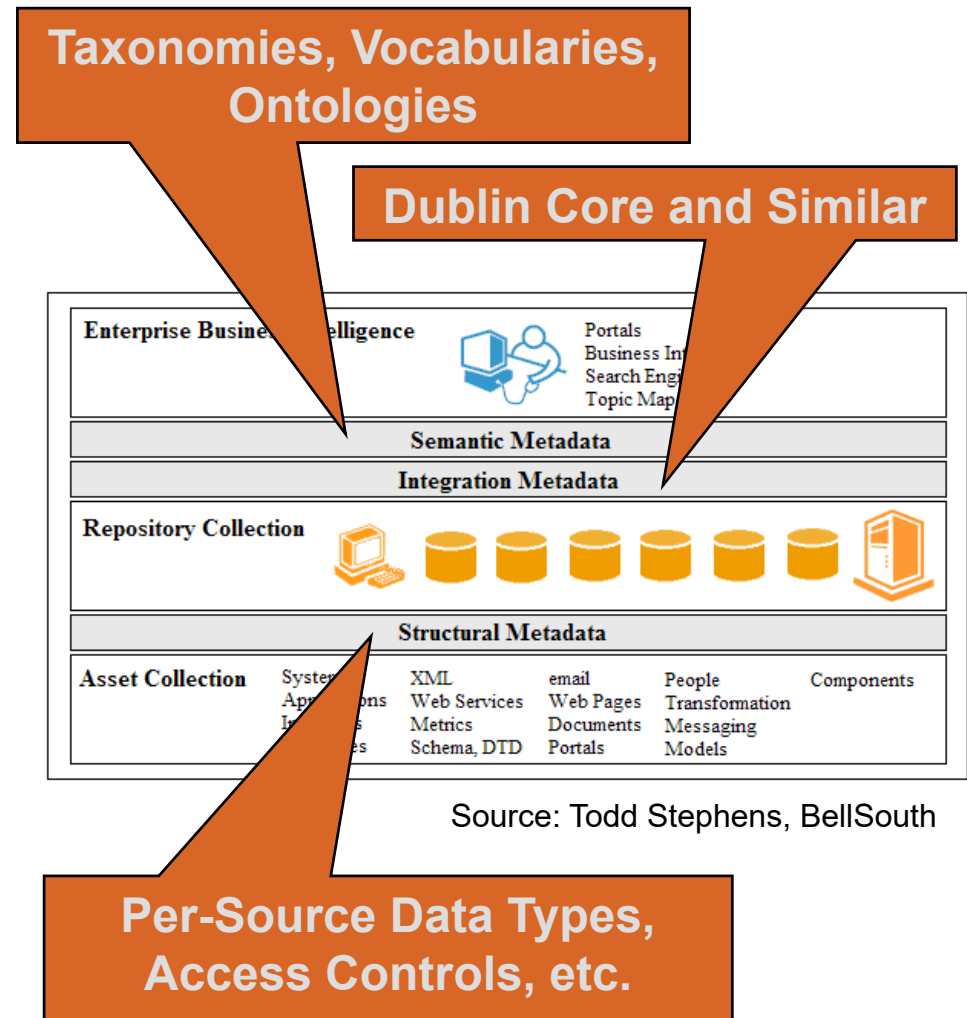
Encodings

None



Sources

- Dublin Core a de-facto standard across many other systems and standards
 - RSS (1.0), OAI
 - Inside organizations – portals, CMS, ...
- Mapping to DC elements from most existing schemes is simple
 - Beware of force-fits
- Why will metadata already exist?
 - Because of search projects, portal integration projects, etc. that are creating it or standardizing a mapping.



Example Source – OAI Harvesting

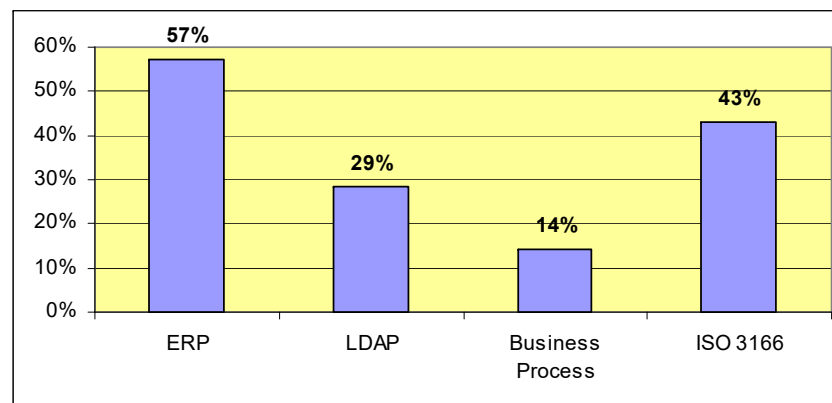
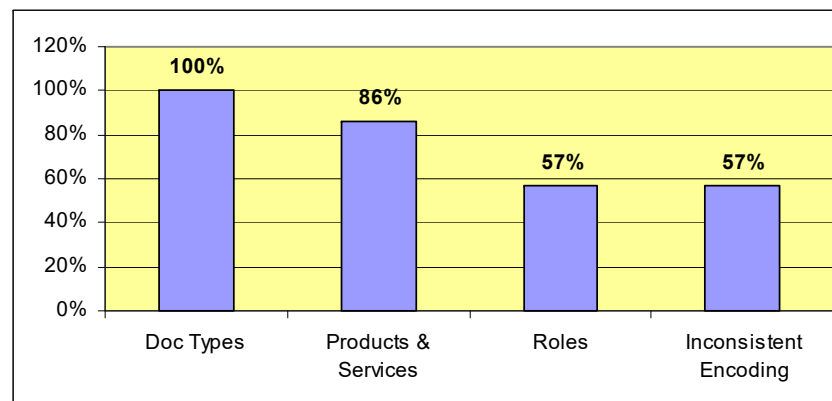
- Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
- Requires a simple Dublin Core format, allows more elaborate formats
 - Tends to get a lot of force-fits into simple DC
- **RSS will be a major source of metadata**
 - RSS 1.0 uses RDF, DC, and has a clearly-stated extension mechanism
 - If you care, bug the Atom WG now.

```
<metadata>
<oai_dc:dc
  xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xsi:schemaLocation=
    "http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
<dc:title>NASTRAN finite element idealization study</dc:title>
<dc:creator>Case, W. R. (NASA Goddard Space Flight Center,
Greenbelt, MD, United States)</dc:creator>
<dc:creator>Mason, J. B. (NASA Goddard Space Flight Center,
Greenbelt, MD, United States)</dc:creator>
<dc:date>1977</dc:date>
<dc:subject>39 - STRUCTURAL MECHANICS</dc:subject>
<dc:relation> - A03 - </dc:relation>
<dc:description>The investigation of the effects of variations of
mesh refinement and mesh pattern were conducted using a basic
rectangular mesh pattern. When employing the constant strain
TRMEM element, the basic rectangular pattern was subdivided into
triangles. This subdivision employs two different triangular patterns ...
</dc:description>
</oai_dc:dc>
</metadata>
```

Source: NASA NTRS OAI Server

Extending the Dublin Core

- Recent study of corporate use of Dublin Core
 - 100% used a custom list of document types
 - 88% added a 'products & services' field of some type
 - 67% added roles and permissions information
- Sources for values
 - 57% used ERP system
 - 43% used ISO locations
 - 29% validated names and roles against LDAP
- Rare to use all DC elements
 - Contributor, Source, ...



Source: Guidance information for the deployment of Dublin Core metadata in corporate environments: CEN Working Agreement (Jan 2005)

Example – NASA Taxonomy Search Prototype

- Top-level taxonomy
 - 11 major branches called facets
 - *Access Rights, Audiences, Business Purpose, Competencies, Content Types, Industries, Instruments, Locations, Missions & Projects, Organizations, Subject Categories*
 - About ½ map to DC elements
 - XML/RDF format vocabulary files
- Metadata spec
 - Based on Dublin Core
 - Facets, plus Title, Date, Description, Creator, etc.
 - XML/RDF format metadata files
- NASA Taxonomy website
 - nasataxonomy.jpl.nasa.gov

Current Search State

The screenshot displays the NASA Taxonomy Search Prototype interface. At the top left is the NASA logo and the text 'NATIONAL AERONAUTICS AND SPACE ADMINISTRATION'. A search bar is located at the top right. Below the search bar, it indicates '222 items matching' and 'Text contains nuclear propulsion'. The search results are filtered by 'Organization: NASA Centers'. On the left side, there are two sections: 'by Organization' and 'by Subject'. The 'by Organization' section lists various NASA centers with their respective item counts: Ames Research Center (3), Glenn Research Center (116), Jet Propulsion Laboratory (36), Johnson Space Center (6), Langley Research Center (9), and Marshall Space Flight Center (57). The 'by Subject' section lists: Aeronautics (5), Astronautics (142), Chemistry and Materials (14), Engineering (11), Geosciences (5), Physics (21), and Space Sciences (11). The main content area shows a search result for 'Nuclear concepts propulsion' with a description: 'Nuclear thermal and nuclear electric propulsion systems will enable and/or enhance important space exploration missions to the moon and Mars. Current efforts are addressing certain research areas, although NASA and DOE still have much work yet to do. Relative to chemical systems, nuclear thermal propulsion offers the potential of reduced vehicle weight, wider launch windows, and shorter transit times, even without aerobrakes. This would improve crew safety by reducing their exposure to cosmic ra...'. Below the description, there are fields for 'Organization: Glenn Research Center', 'Creator: Miller, Thomas J. (NASA Lewis Research Center, Cleveland, OH, United States)', 'Subject: Spacecraft Propulsion and Power', 'Date: 1993-01 (January)', and 'Collection: NTRS'. A 'Focused technology: Nuclear propulsion' section follows, containing a list of topics covered in a viewgraph form, including nuclear thermal propulsion (NTP), high temperature fuel and materials, hot hydrogen environment, test facilities, safety, environmental impact compliance, concept development, nuclear electric propulsion (NEP), long operational lifetime, high temperature reactors, turbines, and radiators, high fuel burn-up reactor fuels, and designs, efficient, high te... The interface also includes a 'Refine by Text Search' box and a 'Sort by Date | Title' dropdown menu.

Facets, Values, and Counts

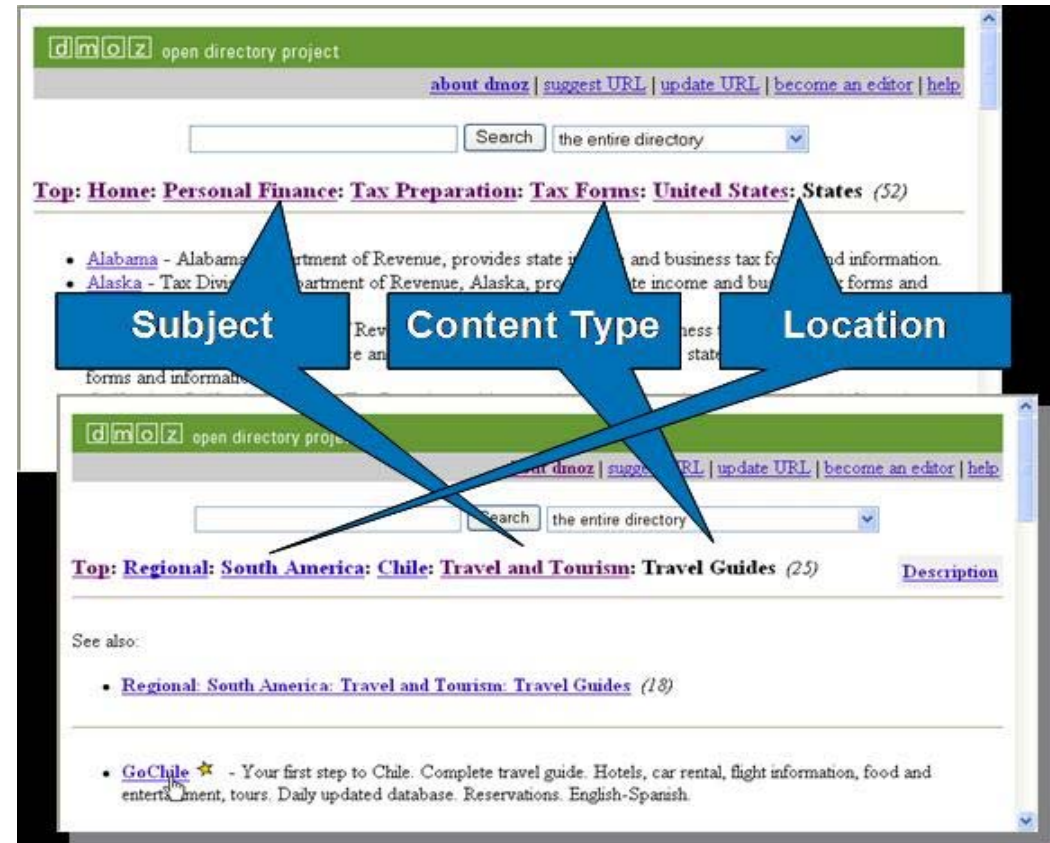
Overview of metadata practices

- Use Dublin Core for basic information
- Extend with custom elements for specific facts
- Use pre-existing, standard, vocabularies as much as possible
 - ISO country codes for locations
 - Product & service info from ERP system
 - Validate author names with LDAP directory
- Design a QC Process
 - Start with an error-correction process, then get more formal on error detection
 - Large-scale ontologies may be valuable in automated error detection



Factor “Subject” into smaller facets

- Size
 - DMOZ tries to organize all web content, has more than 600k categories!
 - Difficulty in navigating, maintaining
 - Hidden facet structure
- “Classification Schemes” vs. “Taxonomies”



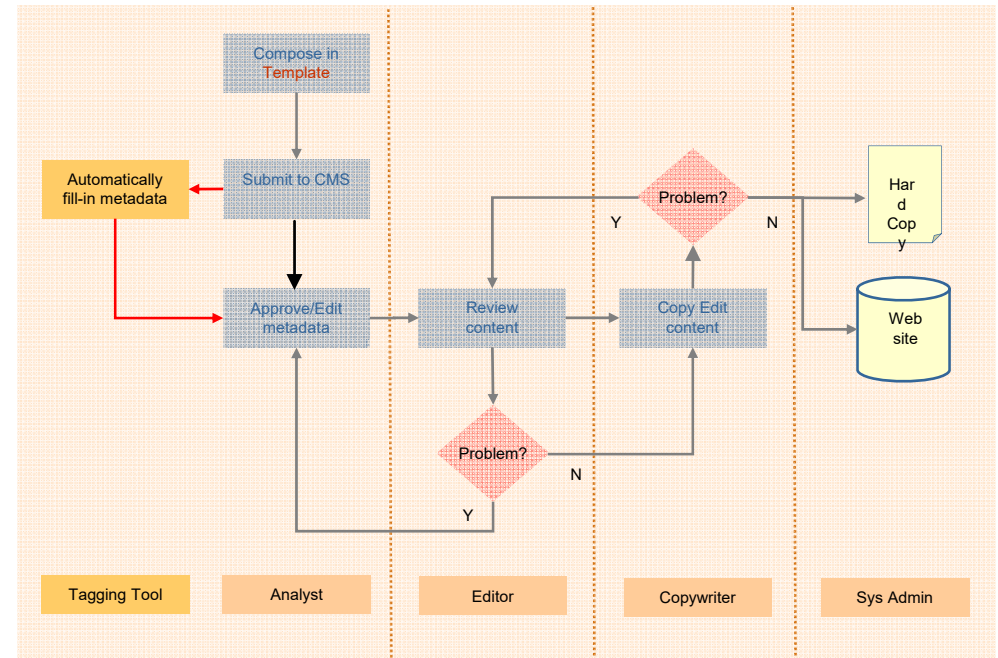
Cheap and Easy Metadata

- Some fields will be constant across a collection.
- In the context of a single collection those kinds of elements add no value, but they add tremendous value when many collections are brought together into one place, and they are cheap to create and validate.



Metadata tagging workflows

- Even 'purely' automatic meta-tagging systems need a manual error correction procedure.
 - Should add a QA sampling mechanism
- Tagging models:
 - Author-generated
 - Central librarians
 - Hybrid – central auto-tagging service, distributed manual review and correction



Sample of 'author-generated' metadata workflow.

Principles

- Basic facets with identified items – people, places, projects, instruments, missions, organizations, ... Note that these are not subjective “subjects”, they are objective “objects”.
- Objective views can be laid on top of the objective facts, but should be in a different namespace so they are clearly distinguishable.
 - For example, labels like “Anarchist” or “Prime Minister” can be applied to the same person at different times (e.g. Nelson Mandela).

