Pushing the Envelope at NASA
The Development and Use of Large Scale Taxonomies

Jayne Dutra
Delphi Information Intelligence
Jet Propulsion Laboratory
California Institute of Technology

April 20, 2005
JPL Today

Parts Catalogues

Electronic Libraries

E-Mail Archives

Financial Data

Engineering Repositories

Problem Reporting

Where do I find it?

April 20, 2005
Knowledge Retrieval

Chances of Finding Needed Information in a Timely Fashion

0%
Taxonomies Reduce Clutter

• Standards are helpful!

• Make it easy for various audiences to find relevant information
  – Provide quick access for NASA Web resources
  – Share knowledge by enabling users to easily find text files, databases and tools
  – Provide search results targeted to user interests
  – Enable the ability to move content through the enterprise to where it is needed most
  – Facilitate Records Retention and Management

• Comply with the eGov Act of 2002
Life Cycle of Electronic Content in the Real Time Organization

Create
- Content
- Assets

Classify
- Logical & Intuitive Filters
- Taxonomy

Discover
- Site Maps
- Search Engines
- NASA Portals
- Content Integration Networks

Finding the right information at the right time to solve the problem at hand

April 20, 2005
So How Did We Start?

- **Content Audit and Analysis**
  - Where is it? How is it published?
  - What is it about? Who cares?

- **Communities Definition**
  - What are the significant knowledge domains?
  - Stakeholders, publishers, consumers

- **Semantic Frameworks**
  - Controlled vocabularies from subject matter experts
  - Conversations with "rocket scientists" (and managers and engineers and accountants, etc, etc)
Taxonomy Basics

What is the NASA Taxonomy?
• A classification scheme meant to encompass all of NASA web content, including internal as well as external material. It is a means for tagging content so it can be used and reused in different contexts.

How to Use the NASA Taxonomy
• This is a generic taxonomy from which specializations can be derived for specific purposes
  – A facet is a branch of the taxonomy
  – Not all facets need to be used in each instance
  – A facet is repeatable
  – The taxonomy is modular and dynamic
Best Practices increase interoperability and extensibility

- Faceted Classification Schema
  - Facets give flexibility and power
  - Modular in nature for easier maintenance
  - Can tag what is appropriate to the use case

- Polyhierarchy
  - Concepts can appear more than once
  - Enables knowledge discovery from multiple viewpoints
  - User-centric organization
NASA Taxonomy Best Practices

• Hierarchical Granularity
  – Different levels of depth depending on attribute set and content
  – The NASA taxonomy is broad in nature by design
  – Integration points allow for mapping of local vocabulary terms back to larger semantic framework
    • Enables schema reconciliation

• Use of Existing Standards
  – Incorporates existing federal and industry terminology standards like NASA AFS, NASA CMS, FEA BRM, NAICS, and IEEE LOM
  – Provides for NASA XML namespace registry (DISA)
  – Complies with metadata standards like Z39.19, ISO 2709, and Dublin Core
But Is It Right??

Test and Validation Phase

- Qualitative validation
  - Confirm stakeholders and communities
  - 71 interviews completed
- Quantitative validation
  - Select and build test collection
  - Stratify automated categorizer
  - Taxonomy test and demonstration - Seamark
- Extend taxonomy value space as needed

This work is iterative

April 20, 2005
### Expressing Strategic Value With Semantic Frameworks

<table>
<thead>
<tr>
<th>Facets</th>
<th>Strategic Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access Requirements</td>
<td>Sensitivity and access control</td>
</tr>
<tr>
<td>Audiences</td>
<td>Who is the content intended for</td>
</tr>
<tr>
<td>Business Purpose</td>
<td>Why the content was created</td>
</tr>
<tr>
<td>Competencies</td>
<td>Relevant field or discipline</td>
</tr>
<tr>
<td>Content Types</td>
<td>The genre of the content</td>
</tr>
<tr>
<td>Industries</td>
<td>External partners &amp; businesses</td>
</tr>
<tr>
<td>Instruments</td>
<td>Flight payloads that yield science</td>
</tr>
<tr>
<td>Locations</td>
<td>Sites where work occurs – on and off Earth</td>
</tr>
<tr>
<td>Missions/Projects</td>
<td>NASA’s lines of business</td>
</tr>
<tr>
<td>Organizations</td>
<td>NASA organizations</td>
</tr>
<tr>
<td>Subject Categories</td>
<td>The topic of the content</td>
</tr>
</tbody>
</table>
NASA Taxonomy Website

Background and training materials

Links to Controlled Vocabularies

Link to Metadata Specification

Link to XML DTDs and RDFs

NASA Taxonomy - Top Level Facets

- Access Security Requirements
- Audiences
- Business Purpose
- Competencies
- Content Types
- Industries
- Instruments
- Locations
- Missions and Projects
- Organizations
- Subject Categories

What is the NASA taxonomy?

The NASA taxonomy is a controlled vocabulary that is designed to populate the NASA metadata core specification.

It is also a means of tagging NASA content so that it can be used and reused in many different contexts.

Tips on using the NASA taxonomy:

http://nasataxonomy.jpl.nasa.gov

Contact the NASA Curator
NASA Official: Jayne Dutra
Last Updated: May 25, 2004

April 20, 2005
• Content Types
  - Announcements
  - Press Kits
  - Press Releases
  - Articles, Notes, and Papers
  - Calendars and Schedules
  - Agendas
  - Case Studies
  - Catalogs and Databases
  - Correspondence
  - e-Mails
  - Memos
  - Databases
  - Bibliographic Databases
  - Image Databases
  - Designs and Specifications
  - Configuration Controls
  - Notebooks
  - Quality Control
  - Requirements
  - Drawings
  - Educational Materials
  - Activity Guides
  - Educational Toys
  - Educator's Guides

Configuration Controls

**Broader Terms:**
- Designs and Specifications

**Scope Note:**
Records of changes to documentation or hardware, including engineering change requests and waivers.

**Term Number:**
52
Using the Taxonomy

- NASA Taxonomy provides controlled vocabularies used to populate elements of more complex metadata schema such as the Dublin Core (www.dublincore.org)

- The taxonomy facets map to the DC metadata tags

<table>
<thead>
<tr>
<th>Field</th>
<th>Name space</th>
<th>Definition</th>
<th>Data Type or Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>dc:type</td>
<td>The nature or genre of the content of the resource</td>
<td>Values come from NASA taxonomy facet: Content Types</td>
</tr>
</tbody>
</table>
Building Bridges

Dublin Core Mapping and XML Schema Development

• Dublin Core metadata mapping where appropriate
• Created any necessary NASA specific tags
  – Some datasets unique to the Agency
• Developed XML schema from metadata
  – RDF Files enable easy reuse for developers
• Next Steps:
  – Educate and train publishing communities
## Selected and Built Test Collection

<table>
<thead>
<tr>
<th>Collection</th>
<th>Source URL</th>
<th>No of Docs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lessons Learned Database</strong></td>
<td><a href="http://llis.nasa.gov">http://llis.nasa.gov</a></td>
<td>1,370</td>
</tr>
<tr>
<td><strong>SIRTF (Space Infrared Telescope Facility) Project Library</strong></td>
<td><a href="http://sirtifweb.jpl.nasa.gov">http://sirtifweb.jpl.nasa.gov</a></td>
<td>4,054</td>
</tr>
</tbody>
</table>

[http://www.siderean.com/nasa/nasademo.jsp](http://www.siderean.com/nasa/nasademo.jsp)

Demo using Seamark from Siderean Software
Built Demonstration of Taxonomy Value in Search and Navigation

...that provides common access framework across test collections

April 20, 2005
Taxonomy Implementations

- Taxonomy stewardship
  - Governance
  - Maintenance, versioning
  - Education and training
  - Facilitate standard adoption process

- Apply in public and internal portals, applications and repositories
  - DDM, PDMS, CMS, other systems
  - Search integration
    - Faceted search and navigation
  - Content integration networks for real time delivery
Thanks for Your Time!

Jayne.E.Dutra@jpl.nasa.gov
California Institute of Technology

And remember:

Just say NO!
White Papers

- **White Paper: Taxonomy Development With NASA, Dutra and Busch, 2003**

- **White Paper: Implementing the NASA Taxonomy Through Service Oriented Architectures, Dutra and Xiao, 2/2004**
Applying a larger semantic framework to local environments to create an integrated, nested information architecture
JPL Today

Parts Catalogues

Electronic Libraries

E-Mail Archives

Financial Data

Engineering Repositories

Problem Reporting

Where do I find it?

April 20, 2005
What’s Needed At The Center Level

- A **unified view** of the project information space
  - Abstracted away from any particular technology or repository
  - Able to handle “harmonization” between repositories according to a larger point of view

- Common data models within and across JPL knowledge domains
  - Common data elements (i.e. JPL Taxonomy and definitions)
  - Relationships between elements
  - Metadata mappings from local models to domain models

- Semantic models that are part of a cohesive lab wide interoperable information architecture
Unified Project Search: Integrating JPL Engineering Repositories

Case Study Goal: Allow Cassini flight project operations teams to match anomalous behavior from spacecraft to engineering design specifications for problem resolution.

1. Characterize targeted databases/repositories
   Problem Failure Reporting System, Electronic Libraries, PDMS, Risk Management DB, etc.
2. Create RDF from data architectures
3. Queries identify fields of interest using semantic properties and return **integrated** result sets
Cassini Sample of Unified Search

Collections: Problem Failure Reporting System and the Cassini Electronic Library

Not a common metadata schema

**PFR:**
- Project Name
- Anomaly Type
- Subsystem
- Report Status
- Date

**CEL:**
- Project Name
- Content Type
- System
- Project level
- Responsible Team/WBS
- Date
NASA Taxonomy Transitioned to a JPL Taxonomy

Content Types
- Designs and Specifications
- Quality Control

JPL Taxonomy
Terms present in the CEL and PFR
- Problem Failure Report
- Incident Surprise Anomaly
- Corrective Action Notice

Re-Combined through RDF
Cassini Data Rationalization

Connecting heterogeneous collections: PFR System & the Cassini Electronic Library
Mapping fields to each other using semantic hierarchies.

Search and Browse the catalogue by:

- Project Name
- Content Type
- System
- Subsystem
- Responsible Team/WBS
- Date
- Collection

Next terms mapped to each other using an existing spacecraft ontology
Final Results
of Data Harmonization

• A system whereby the user can browse all
documents relating to the Cassini camera
and its subsystem independent of any
particular repository’s search engine.

• Harmonization achieved by mapping terms
to a common vocabulary (the Taxonomy)

• Could browse by:
  – System, Sub-system
  – Instrument
  – Content Type – PFRs, ECR’s, Designs Specs, etc.
  – WBS or Responsible Team
  – Date
Achieving the Vision

• Leverage what projects produce in the normal course of their business
  – WBS lists
  – Document trees
  – Document matrices

• There are many un-mined sources for semantic processing

• What schema already exist in your organization?