NASA Technical White Paper
Enabling Knowledge Discovery: Taxonomy Development for NASA

Jayne Dutra,
Jet Propulsion Laboratory,
California Institute of Technology
and
Joseph Busch,
Taxonomy Strategies

January 7, 2003

Table of Contents

Executive Summary

Taxonomy development provides a framework for authors and service providers working in the Internet age. It allows them to model their content so that it is useful in a larger context such as a target for the federal search engine (FirstGov). Consistently modeled content makes it possible for engineers and scientists as well as the public to find and reuse content, rather than recreate it or make do without it. Adopting frameworks that enable content reuse provide a return on investment (ROI) for the time and effort spent to originally produce the material, or to aggregate it in a portal or other content publishing applications.

The 2002 E-Government Act specifically calls for the development of "standards and guidelines to categorize Federal Government electronic information". The taxonomy work that has been completed to date puts NASA clearly ahead of other federal agencies. This white paper provides the background for why it is important to take the next steps with the NASA Taxonomy including test and validation, XML schema development, integration with the FirstGov federal search engine, the OneNASA portal and its supporting Web content management system.

Background

The National Aeronautics and Space Administration (NASA) is composed of 12 separate Centers that engage in a variety of technological activities with product lines that reflect their specialized work. Technical communities across the Agency utilize highly evolved engineering and scientific vocabularies that reflect the nature of their disciplines. These vocabularies are often arcane and not easily translated by personnel outside a particular community of practice. This situation has tended to fragment information produced by NASA personnel.

Centers also publish material with a wide variety of uses by different communities. Some examples include: spacecraft design and engineering documents, aeronautic testing results, scientific data analysis, administrative data and financial information. NASA is also tasked to share its knowledge with a wider public audience comprised of students, teachers, researchers and the media. The wide variety of constituencies and the amazing breadth of content types calls for a new approach in the management of NASA's intellectual properties.

As the Internet grew in the nineties, more and more information moved to an online format. In 1998 Congress passed the Government Paperwork Elimination Act (GPEA), which requires government agencies to reduce or eliminate paper wherever possible. This move to a new electronic environment is causing NASA to rethink how it identifies, organizes, and retrieves its content assets. Documents are no longer stored as hard copies in filing cabinets or physical archives. Instead, we are beginning to live in a world where knowledge is described in "content chunks" that are held in electronic document management systems.

Unfortunately, the Agency has not yet adopted common standards for handling its electronic assets. Centers use varying solutions for identifying and storing electronic documents that have been independently developed at each facility. The uneven way that these systems have evolved hampers the ability of users to find relevant information, especially when searching a topic that spans work at multiple Centers. The resulting situation is that NASA does not know what knowledge it possesses or where it resides. While knowledge workers need unified, universal access to information across the Agency, they only want the specific information that actually solves their problem at hand. Finding exactly and only the content they need is a problem that has not yet been solved by the Agency.

A similar situation exists in NASA's Web space. Web sites for specific missions or enterprises have been mounted without much thought given to the larger venue where users access sites across a number of Centers. Material is often redundant or out of date. Global navigation is non-existent, confusing users who often times abandon NASA's sites for the better presentation of news sites such as CNN or Space.com. Comprehensive and accurate searching is difficult if not impossible to achieve on NASA sites due to inconsistent tagging and different publishing conventions among NASA Web masters.

These new challenges make it necessary to revisit the processes and technologies used to catalogue NASA's electronic content. The logical place to do so is by specifying a common labeling framework within which material can be consistently tagged. The result will allow knowledge workers to effectively organize, query, use and re-use Agency-wide content.

Concept

The fundamental goal behind building and adopting an agency-wide NASA Taxonomy is to develop a consistent methodology for handling NASA's electronic content. Documents are described with a standard classification scheme that follows a predefined hierarchy. This allows the user to discover and see correlations between subject areas as illustrated in figure 1, a Yahoo search on the word "trial". It also allows the search engine to retrieve information with more precision and relevancy. Every time an engineer or scientists finds and reuses a content chunk, the return on investment (ROI) for the work that originally produced the content chunk increases. This cycle of reuse directly impacts the Agency's bottom line. It also pushes the pace of development forward at a greater rate as teams build on previous work instead of "reinventing the wheel".

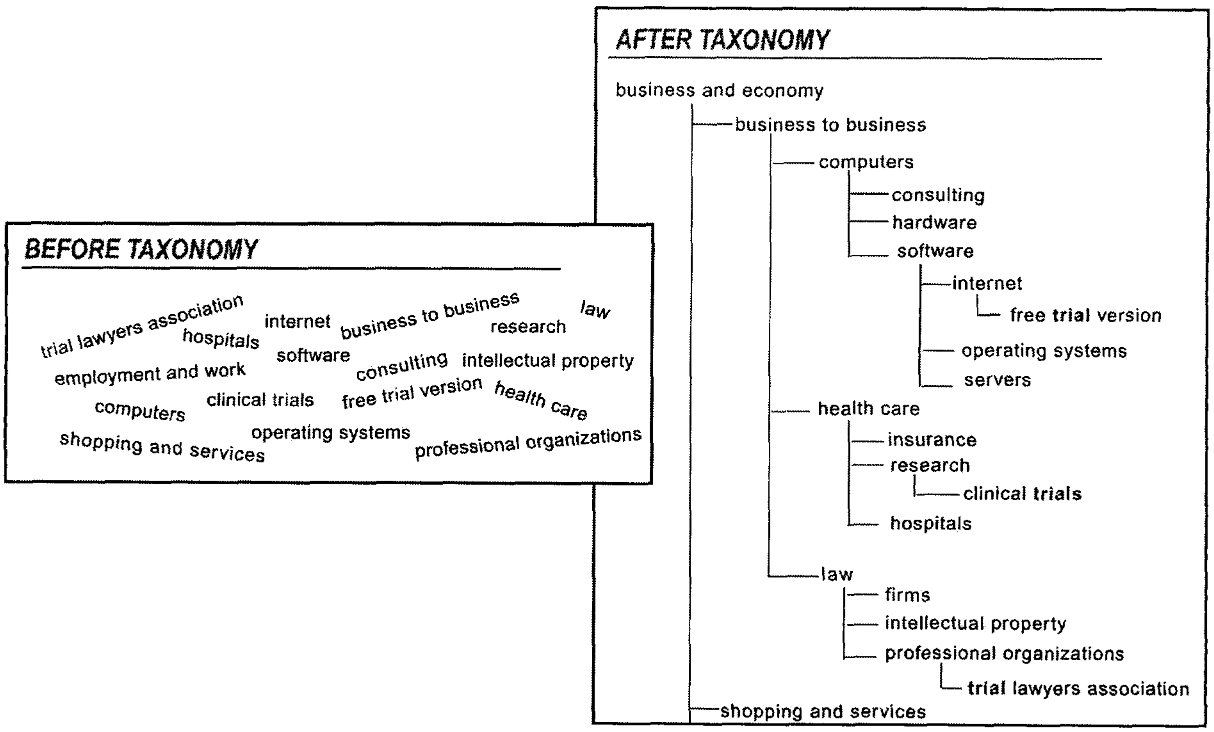

Figure 1 - Yahoo search results for "trial." From: *Taxonomy at a glance.* http://enterprise.yahoo.com/portal/services/taxonomy/taxonomy1.pdf. Last checked Jan. 6, 2003

Taxonomies contain predefined hierarchies of descriptors for marking content chunks. Taxonomies are composed of discrete branches which are also known as *facets*. Facets are made up of hierarchically organized lists of attribute values for use in consistently labeling content components with repeatable attributes. Facets allow taxonomies to be multi-dimensional, which accommodates a wider range of content types. Taxonomies that are designed to be modular and extensible will be the most robust.

Design Approach and Methodology

Taxonomies are best developed from the top down with the help of subject domain experts in the range of fields covering NASA's core work. Development is implemented as an iterative process that identifies and enables content owners while engaging stakeholders. A taxonomy is a means to an end, and not the end in itself. A good taxonomy is not perfect, but it does improve search and navigation immediately once it is put in place.

The following table provides a broad outline of the taxonomy development process and timeline used at the Jet Propulsion Laboratory (JPL) for the development of the NASA Beta Taxonomy. This taxonomy was generated with the consulting help of Joseph Busch from Taxonomy Strategies, a company that specializes in taxonomy, meta data and XML applications for large institutions.

1. **Phase 1 – Scope taxonomy**
   a. Start up and introduction
   b. Interview stakeholders and subject matter experts
   c. Collect exemplary documents and existing controlled vocabularies
   d. Analyze data collected
2. **Phase 2 – Build Taxonomy**
   a. Develop alpha taxonomy
   b. Present results to stakeholder community and gather feedback
   c. Review, revise and reach consensus on alpha taxonomy
   d. Develop beta taxonomy
   e. Present beta taxonomy to stakeholder community
   f. Review, revise and reach consensus on alpha taxonomy

NASA Taxonomy Development Roadmap

Phase 1: Develop Alpha Taxonomy

The initial taxonomy that will seed the collaborative effort is specified in this task. Alpha taxonomies are typically a top-level outline 2 to 3 levels deep. Top-levels are usually the most stable component of a taxonomy. For this reason, the effort to reach organizational consensus is focused on the overall outline, while validation of sub-trees and leaf nodes is delegated to subject matter specialists.

The alpha taxonomy may be generated by one or more of the following mechanisms:

Inspiration	Viewpoint of an individual consultant or project participant.
Induction	Analysis of document collections using automated tools.
Deduction	Application of an existing standard, approach, or model.
Synthesis	Combination of approaches.

Alpha taxonomies are typically generated via synthesis, using a combination of these techniques.

The NASA Taxonomy was developed based on semi-structured interviews with Subject Matter Experts (SME's) and collection and analysis of content culled from existing controlled vocabularies, exemplary web sites, and other resources that SME's supplied to the taxonomy team.

Critique and revise alpha taxonomy. Alpha taxonomies are typically reviewed in outline format. Generally, each participant independently communicates their critique and comments in writing. In some cases a more structured process utilizing a questionnaire may be needed to effectively reach consensus.

Stakeholders responded to the alpha NASA Taxonomy in group meetings, via email, and in one-on-one conversations.

Reach consensus on alpha taxonomy. Iterate until consensus is reached.

Phase 2: Develop Beta Taxonomy

Filling out the agreed upon taxonomy outline typically makes use of standard and internal resources discovered during the scoping research and analysis, rather than original material. Building on the solid foundation of the alpha top-level outline, filling out the taxonomy is a straight forward process. A comprehensive taxonomy consists of approximately 1500 terms and as many as 6000 nodes.[1]

The beta NASA Taxonomy was prepared based on stakeholder feedback and taxonomy best practices. A public version of the Beta taxonomy can be reviewed at: http://eis.jpl.nasa.gov/~jedutra/NASA_Taxonomy_ver3/Index.htm.

Review and revise beta taxonomy. Beta taxonomies may be reviewed in whole by project participants, or divided up into trees and reviewed by specialized subject matter experts in parts. In some cases, it is advisable to recruit third party experts to validate the complete taxonomy. As with the alpha taxonomy review, participants independently provide critiques and comments in writing.

Reach consensus on beta taxonomy. Iterate until consensus is reached.

Best Practices

The development of the NASA Taxonomy is based on industry best practices in the library science and information architecture communities.

Hierarchical Granularity. The taxonomy is designed to provide as much depth or hierarchical granularity in the classification as the content requires. Because NASA's content includes highly technical subject matter, this allows authors to tag their material with precision, which also enables better search mechanisms.

Polyhierarchy. The taxonomy allows the same concept to reappear multiple times in the scheme. Because the same concept can then have multiple parents, navigational pathways are built in that facilitate a search from numerous and different approach points. Instead of knowing exactly the right term to search on, a user can come from his or her own individual viewpoint and still locate the pertinent information.

---

[1] Concepts frequently occur in multiple hierarchies, i.e., a term has multiple parents. For example, the concept "insurance" can occur in a Business and Finance taxonomy in hierarchies related to legal, research, regulation, industry, institutions, finance, as well as commerce. The assumption is that on average, a concept will occur in 4 hierarchical locations.

[2] Concepts frequently occur in multiple hierarchies, i.e., a term has multiple parents. For example, the concept "insurance" occurs in our Business and Finance taxonomy in hierarchies related to legal, research, regulation, industry, institutions, finance, as well as commerce. The assumption is that on average, a concept will occur in 4 hierarchical positions.

Mapping Aliases. To add more richness to the labeling scheme, abbreviations and alternate terms are mapped to the taxonomy. By planning for acronyms and synonyms early on, the taxonomy becomes more accessible to users that possess a deeper grasp of any one topic area.

Existing Standards. Throughout the development period, effort was made to adopt categories for standard genre and document types in the Information facet of the taxonomy so that users could start with a common understanding of what document frameworks they might be looking for.

The facets that resulted from the interviews with NASA Subject Matter Experts conducted in Phase 1 of the development task were standardized to reflect NASA's work. The table below shows some samples.

Facet Name	Correlation to NASA Business Goals
Disciplines	NASA's technical specialties (engineering, scientific, etc)
Functions	Business records and record management
Industries	NASA's partners and other entities that we do business with
Locations	Sites on Earth and off Earth
Organizations	NASA affiliations and organizations it is a part of
Projects	NASA missions, projects, product lines, etc

Table 3 – NASA Taxonomy Facet Samples

Significantly, the development team determined that it is most useful to reuse existing standards and vocabulary sources that have already been adopted by various technical communities. Therefore, the Association for Computing Machinery (ACM) specification is used for the Computer Science facet, the Agency Filing Scheme (AFS) spec managed by the NASA CIO's office is used for the Functions facet, the Learning Object Metadata (LOM) standard is used for Educational Roles, etc. This reuse of professional standards ensures that users who are familiar with the standards of their own field will be comfortable with the standards being applied to documents they author. It also means that the resulting taxonomy is ready to be plugged in to a larger federal taxonomy when the time comes to join with other agencies. The increased interoperability adds real value and longevity to the work produced.

Applications and Benefits

The NASA Taxonomy addresses the Data Reference Model layer of the federal architecture as laid out by the Federal Architecture Solutions Group in the U.S. Office of Management and Budget. A table delineating the various component layers is shown below. The value from the standardization of NASA's data extends not only throughout the Agency itself, but to a wider application in the overall federal information landscape.

Federal Enterprise Architecture

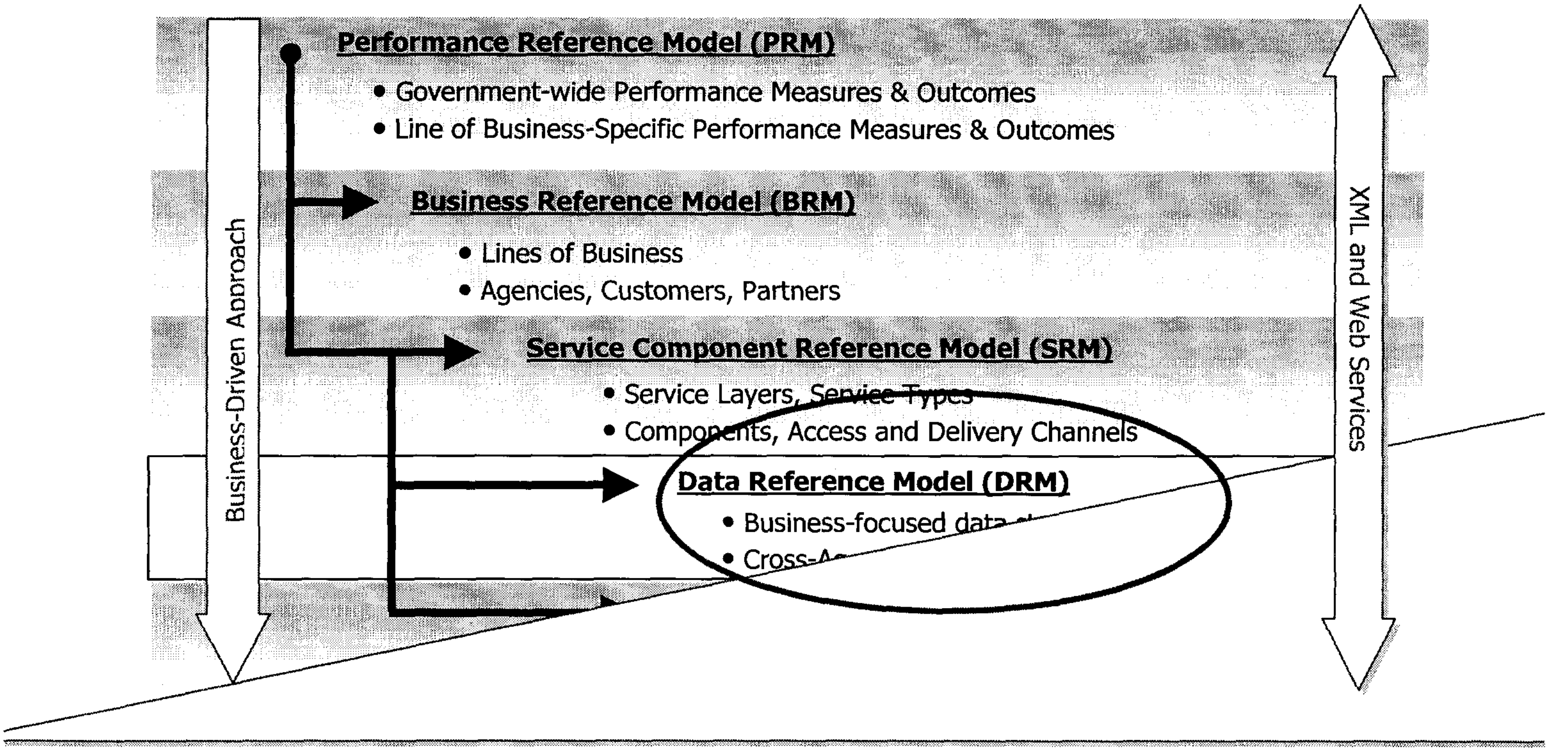

Figure 2- Federal Enterprise Architecture, XML and Web Services Across the Federal Government. Robert Haycock, Program Manager, FEA-PMO. Presentation at *XML 2002*, Baltimore, MD, Dec. 10, 2002.

Once a taxonomy is proved to be viable through test and validation exercises, many uses are possible for augmenting the application layer of NASA or federal systems. The underlying taxonomy adds flexibility and power to search, metadata development, XML applications for content management systems (management of material for large Web sites), document and data management systems, Web services applications and navigational applications such as site maps and Web directories. Let's examine two scenarios from these cases to better understand the value derived from the addition of a taxonomy layer.

Application Example #1: Adding Intelligence to NASA Search

Conventional Web search engines send out "spiders" to crawl and catalogue organizational Web space. Most of these search engines can only index whatever material is placed there by Web publishers. Many times, these publishers may not be familiar with the hidden benefits of metadata or the use of consistently applied controlled vocabularies. Because metadata tags are not displayed on the browser, they are often forgotten or underutilized. Without a stable data construct to consult, search engines return page after page of results that are not truly relevant to a user's query.

What users really need are results that fit the context of what they're asking. They need the search engine to look beyond the ambiguity of natural language and somehow "know" which knowledge fragment they require to solve the problem they are facing at that moment. When documents and Web pages are tagged in a descriptive manner that reflects an ordered hierarchy, they are enriched in a number of ways. Metadata tags carry markers that allow the search engine to filter results based on what branch of the taxonomy they reside in. Right away, the user can determine if a result set is appropriate or useful. For example, a search for "Odyssey" could turn up both the NASA mission to Mars as well as Homer's epic. Which return does the user really want? Metadata tags can contain content descriptions that assist the user in "zeroing in" on the right content. Metadata tags can tell the user what version of the document they are examining or if it has been approved for publication. They can tell where something was published and what audience it is intended for.

Taxonomies enable more than metadata tagging for search engines. They also provide a way to map alternate terms to root words that might include abbreviations, synonyms or other aliases as illustrated in figure 3, screenshots from MSWeb Best Bets. It might be highly desirable for a search on El Nino information to return the TOPEX/Poseidon home page for example since this NASA mission studies the effects of the El Nino phenomenon. This could be done through the mapping of subject classification schema to alternate phrases in the taxonomy. NASA abounds in acronyms–many of our mission names are shortened and then referred to by their acronyms. This works well for internal teams with insider knowledge, but can pose obstacles for members of the public who are not well versed in this specialized vocabulary. Creating a back end thesaurus customized for use with the NASA search engine that specifically encompasses these types of acronyms would be a tremendous help in assisting students and others to find information on NASA Web sites.

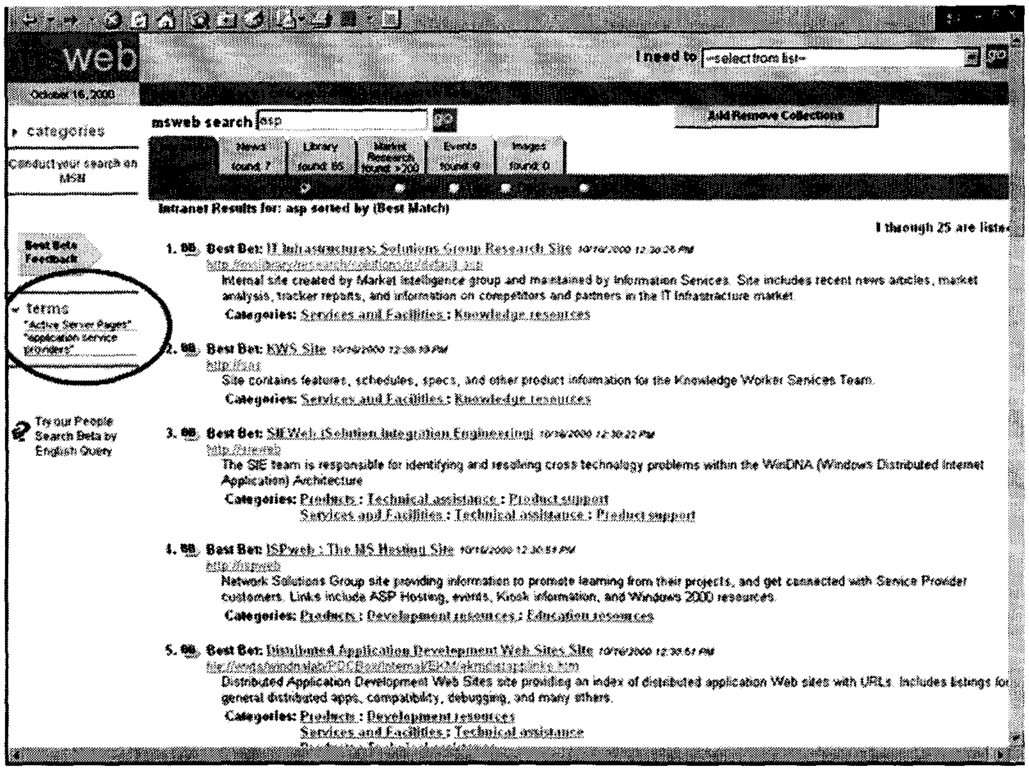

Figure 3 - MSWeb Best Bets automatically expands query with aliases and suggests related terms. Vivian Bliss presentation at *Information Architecture 2000*, La Jolla, CA, Oct. 27, 2000. http://argus-acia.com/acia_event/bliss_session.html. Last checked Jan. 6, 2003.

Application Example #2: Taxonomy and Web Publishing on the NASA Portal

Recently, a task was undertaken to create a new public Web home page for NASA. This home page is conceived not as a static HTML file as was common in the past, but rather as a portal that unites many different content streams and is customizable by visitors to reflect their particular interests and knowledge levels. The content flowing through this portal is dynamic–it changes every day. This level of change is new for NASA Web masters. It is especially challenging because they are not necessarily aware of what other authors may be posting at different Centers. In order to deal with this altered situation and also to leverage the existing content on sites scattered across NASA Web space, new methods of authoring and publishing online must be developed to coordinate the results. These should involve implementing content management systems based on XML schema, namespaces, and metadata applications.

In addition, the editors of the NASA portal have begun to organize the way material is presented to their users. They may create several different online "departments", much like a magazine or newspaper. They may debut their new site with one set of subject categories defined especially for their editorial intent. Once their initial categories are defined, they will turn to the content management programmers to have them "hardwire" the categories to the portal framework. They will ask Web authors to add metadata tags and other content descriptors to their material before it comes into the portal. Web designers will build navigation bars and site maps that match the category definitions. Much time and effort will go into modeling all of the content to these categories.

But what if three months after the portal's unveiling, the editors begin to learn more about the needs and requirements of their audience and want to change the way that content is presented? They may realize that certain areas of the site are underutilized. Or there might be other areas of the site where users are demanding more in-depth content. NASA may uncover some significant breakthrough technology or new scientific knowledge that alters the way we think about space exploration or life on Earth. There are many, many possible reasons why the editors may decide to change the site. This is the norm in Web site design today and its eventuality should be planned for.

Unfortunately, if the site does not have the underlying layer of a neutral labeling framework, the amount of resources to revise the portal will be tremendous. New funding will have to be allocated. Application programming will have to be redone. Existing material will have to be re-tagged and authors will spend time undoing previous work instead of writing new content. All of this slows down the time-to-market for Web material in an environment where the shelf life of information has been continuously shortened. Web users today expect the latest in informational output and certainly NASA wants to demonstrate high quality in content delivery services.

If, on the other hand, an underlying taxonomic framework had been put in place first, the content would be identified in a way that allows site editors and designers the freedom to re-configure their portal without the headaches and increased cost cited above. The taxonomy acts as a neutral labeling mechanism to create building blocks that may be put together in any way site owners decide. Metadata tags and XML schema, which are derived from the taxonomy, separate the content descriptions from the presentation format determined by portal editors. The flexibility and robustness of the content is enhanced and portal editors are given a new level of options that was not possible with the flat file organizational schema tied to the presentation and architecture of one site.

Taxonomy as a Foundation for XML Applications

XML schemas are semantic data models expressed in XML. They are composed of elements that define structured information hierarchies and express the meaning of content. XML elements enable intelligent content processing and allow filtering of relevant and irrelevant materials based on the metadata and associated business rules they contain. Metadata elements provide the underlying structure that describes the subject classification breakdown as determined by the source taxonomy. In this way, the taxonomy is leveraged to form the foundation for an information network above any kind of resources or data objects it refers to.

Taxonomies can overlap information spaces and allow them to interrelate. This higher meta level of taxonomy formation is called an ontology. An ontology is defined as "an explicit specification of a shared conceptualization." [3] A conceptualization consists of relevant concepts of a domain, the relationships between these concepts and agreed upon knowledge about these concepts by a group. A formal ontology enforces well-defined semantics on a conceptualization, which can be described through XML elements.

Ontologies can be used as interchange formats, enabling mediation between systems in a Web Services model. When implemented with controlled vocabularies and taxonomic underpinnings, ontologies enhance reusability, search results, reliability, and knowledge acquisition. Ontologies and topic maps can allow us to catalogue what we know and what we don't know, helping to shape our future research efforts as an Agency.

The E-Government Act of 2002 and Challenges for NASA Publishers

Today NASA knowledge architects are challenged by a heightened awareness at the federal level of the importance of establishing new frameworks for information technology standards. The speed and quality of taxonomy development and adoption puts NASA in a leadership position for compliance to the requirements of the new law.

In 2002, the U.S.Congress passed the E-Government Act. This Act specifically calls for the development of *"standards and guidelines to categorize Federal Government electronic information".* In addition, Section 207 of this Act states that its purpose is to *"improve the methods by which Government information, including information on the Internet is organized, preserved, and made accessible to the public in a way that is searchable electronically and interoperable across agencies..."*

Item F of Section 207 calls out requirements for federal Web sites including minimum agency goals to assist public users to navigate agency Web sites, in particular focusing on *"the speed of retrieval of search results, the relevance of the results, and tools to aggregate and disaggregate data..."* [4] All of these goals are enriched by a robust taxonomy.

The taxonomy work that has been completed to date puts NASA clearly ahead of other federal agencies. As has been described above, the maximizing of commercial standards, the accent on federal interoperability in a wider information architecture and the ability to impact search and data aggregation are all well aligned with the E-Government Act

Next steps for the NASA Taxonomy include the test and validation phase of the beta version and its mapping to a set of Dublin Core metadata tags as shown in Table 4. An initial pass on the metadata mapping has been completed, however the testing still needs to be done. Once the taxonomy is validated, activity can begin on integrating it into the NASA content management system underlying the new OneNASA portal. This will most likely involve XML schema development which should be done in concert with current work at the federal level to ensure reuse and interoperability. In addition, work can begin on a thesaurus for the Agency search engine.

Challenges for the future entail the support and ongoing maintenance of the Agency's taxonomy. Since taxonomies reflect the current state of knowledge by an organization, they are, by their very nature, organic. As the collective knowledge of a group grows and changes, so should the taxonomy. A key recommendation for this taxonomy project is to build a community of NASA publishers who are committed to ensuring that the taxonomy stays current on a long term basis.

---

[3] Tom Gruber, Stanford University

[4] http://xml.gov/documents/completed/eGovXML.htm

Dublin Core Elements	Definition	NASA Taxonomy Mapping
Creator	Content maker.	dc:creator dc:creator.employee dc:creator.organization
Subject	Content topic.	dc.subject.organization dc:subject.missionsProjects dc:subject.disciplines
Publisher	Publisher of this manifestation.	dc:publisher.organization
Contributor	Content contributor.	dc:contributor dc:contributor.employee dc:contributor.organization
Type	Genre.	dc:type.information
Coverage	Space, period, date, jurisdiction, etc.	dc:coverage.locations dc:coverage.chronology
Audience	Content audience.	dc:TERM.audience
Non-DC	NASA missions and projects.	nasa:missionsProjects
Non-DC	Business functions.	nasa:functions
Non-DC	Technical specialties.	nasa:disciplines
Non-DC	Standard industry categories	naics:industries

Table 4 - Preliminary NASA Taxonomy mapping to Dublin Core

The Future of Taxonomy Development

It is clear from the plans of the Federal Enterprise Architecture Program Office of the OMB that the federal government is determined to proceed down a road that unifies as many of its service providers as possible. The technologies that it is relying on to achieve this transformation of information management lies in the areas of XML applications and Web Services. Both of these technologies in turn derive their integrity and extensibility from the underlying foundation layer of viable taxonomies.

The NASA Taxonomy has not yet been fully completed. Below is a roadmap in outline form of proposed activities and deliverables for a final product ready for integration into various NASA applications[5].

- Development of a targeted content set used for testing
- Testing and validation – several iterative cycles
- Consensus by the NASA publishing community
- Derivative meta data tags created and adopted
- XML modeling and schema development
- Integration into Agency Content Management system
- Integration with Agency Search Engine
- Integration with Agency Web Site Registration System
- Integration with FirstGov search taxonomy
- Community formation for long term maintenance and taxonomy extensions

[5] *NASA Web Information Proposal, Dutra and Busch, 2002*

NASA has begun to participate in the Federal XML Working Group. Hopefully, NASA personnel can begin to better understand how other federal agencies are incorporating XML into their electronic services as well as sharing their experience and lessons learned.

Summary

Taxonomy development provides a framework for authors living in the electronic age. It allows them to model their content so that it is useful in a larger context such as a target for the federal search engine (FirstGov).

Knowledge workers spend 26% of their time re-creating existing content they cannot find and only 9% of their time creating new content!

In a study by the IT market analysts IDC[6], the findings show that today's workers spend up to 2.5 hours each day looking for information. Unfortunately, workers find what they are searching for only 40% of the time. In the most disturbing finding of all, because they cannot locate targeted information, workers spend more time recreating existing content than creating new content.

NASA, like all federal agencies, needs to make the best use of workers' time. When an engineer or scientist finds and reuses content, the return on investment (ROI) for the time and effort to originally produce the material increases. The cycle of creation and reuse directly impacts the Agency's operating costs. It also pushes the pace of development forward at a greater rate as teams build on previous work instead of "reinventing the wheel" over and over again.

The business case for maximizing NASA's ability to locate and reuse existing content rests on better management of the intellectual assets of NASA's scientists and engineers. The modernization of the Agency's approach to handling its electronic wealth should become a priority of critical importance today.

6 *The High Cost of Not Finding Information*, 2001, IDC

Acknowledgements

We would like to thank all the individuals who participated in the interviewing process and were kind enough to give feedback to the alpha taxonomy. Their domain knowledge and willingness to support this activity made a crucial difference to the quality of the work.

This work was performed at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.