A NASA Technical White Paper: Implementing the NASA Taxonomy Through Service Oriented Architectures to Promote Knowledge Sharing and Increased Mission Success

Jayne Dutra
Section 366, JPL Knowledge Management Project
Qing Xiao
Section 381, Scientific Data Management

Jet Propulsion Laboratory
California Institute of Technology

JPL D- 27840
Revision 1.0
Draft Version

February 4, 2004

Table of Contents

Introduction

The amount of content produced and published to the Web continues to grow at a brisk pace. According to a recent study by the University of Berkley, there are now 250 megabytes of content for every person on the earth and we are continuing to create new content at the rate of one to two billion gigabytes a year.[1] The sheer amount of content makes finding any one piece of relevant data a daunting and time consuming task. Additionally, the big IT investments of the late nineties encouraged the use of proprietary technologies. These technologies meant that creating linkages between systems required APIs needing hard coded interfaces. The results have been fragmented information spaces in legacy applications that do not fit with current NASA business goals.

Up until the present, NASA employees have had to guess where a particular piece of information might reside in order to query the system and retrieve it. Not only do users need to know where exact types of data are located, but they also need to know the correct key words to enter into the search box. Because there isn't yet a consistent set of controlled vocabularies in use across the Agency, keywords associated with documents can be pulled from anywhere by the author and they might be expressed in highly technical language that is unknown to the worker trying to access content further down the information food chain.

Clearly this is an environment where the chances of finding exact information at the time it is needed are very low. In the NASA 2003 Strategic Plan, one of the Agency's primary goals is "*to develop new capabilities and revolutionary technologies that will change the definition of what is possible*". It goes on to say that *"we will assemble new tools and architectures to provide an intuitive, highly networked engineering design environment that will unleash the creative power of engineers and technologists across the Nation."* In order to achieve such lofty goals, we need to rethink how we are providing information to NASA employees.

Workers today expect to access material in a self service information environment. Because the trend towards content publication on the Web is accelerating, it is apparent that more efficient ways to manage content are critical to any enterprise wanting to be successful in a dynamic information environment. The goal of this paper is to briefly describe new technologies available to us and discuss the strategic value they have for the NASA enterprise with some considerations and suggestions for near term implementations.

Finding Information - Still an Issue

Data repositories across the Agency still reside in isolated environments due to security concerns and legacy system implementations. Current interfaces between these legacy systems are tightly bound by proprietary APIs and there is currently no institutional method in place to integrate searches spanning multiple repositories. In order to retrieve a particular piece of information, one must go out looking for it in many places. This requires an employee to interrupt his flow of work and take the time to mount a search for data needed to complete a particular task.

Studies show that users spend approximately 25% of their time in retrieval activities[2]. Not only do users spend significant time seeking information that they think might exist, but if they are unsuccessful in locating a piece of data, they will frequently recreate the material so they can continue on with their tasks. If the information remains elusive and hard to pinpoint, workers might also contact a librarian or other expert to help them locate data in a mediated search. However, this extra step is also time consuming and most workers just want to download that piece of data and move on.

The latest evolution in information delivery goes beyond individual search engines or browse mechanisms. The next phase of data retrieval involves pushing information to the worker at the

---

[1] M. Strohein, S, Stearns, *Content Management That Fuels the Real Time Enterprise,* Outsell and Inmagic, 2003.

[2] IDC, *The High Cost of Not Finding Information*, 2001.

appropriate moment it is needed, a "just in time" delivery model. Content is said to be made "portable" by implementing an infrastructure architecture that is capable of querying multiple enterprise systems and moving the content dynamically to where it is needed at a precise moment in time. The concept of dynamic content moving about the organization to be retrieved and displayed through various Web applications is becoming known as a "content integration network."

Business Drivers for Changing Models of Content Delivery

As the move to the Web for information retrieval becomes widespread among citizens, federal agencies have become more aware of their need to provide content that is relevant and valuable to them. 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.

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..."* [2] All of these goals are enriched by a robust taxonomy.

When George Bush signed the E-Government Act, he mandated that all federal agencies offer governmental information and services on the Web. In response, the Office of Management and Budget developed a new Federal Enterprise Architecture, which is heavily based on XML Web Services and the notions of consistent data modeling for better content dissemination.[3] The Federal Board of CIOs has created two working groups to support just such development activity–the Federal XML Working Group and the Federal XML Web Services Working Group. Both of these groups are working towards standardization of data models and building an architecture based on common infrastructure components.

Records Management is another business driver for creating robust content infrastructures. The National Archives recently issued a draft version of new guidelines for archiving Web content[4]. This will result in even more Web content being available for users to sift through. As we consider the future of the Agency's use of the Web for mission development, it behooves us to think about archiving content automatically as part of the broader process of content management in order to meet these new guidelines.

Agency Drivers for Changing Models of Delivery

In addition to the events in the federal arena, NASA has recently seen some internal developments key to the implementation of content integration networks. The creation and delivery of a stable NASA taxonomy in spring of FY 04 marks the first time the Agency has adopted a consistent reference model for its content. The release of an Agency taxonomy provides a common semantic framework that developers can build to while being sure that their components integrate into a larger architecture.

Internal and external portals have been rolled out this year for NASA and next steps in their development involve implementing capabilities to collect and display information based on metadata attributes. Project portals are now in use by JPL flight teams and, with the development of an institutional information architecture, their true value as aggregators can be leveraged to automatically discover relevant

---

[3] Federal Enterprise Architecture - http://www.feapmo.gov/fea.asp

[4] . "Endorsement of DoD Electronic Records Management Application" (January 15, 2003) http://www.archives.gov/records_management/policy_and_guidance/bulletin_2003_03.html

information silo'ed across multiple hosts and diverse applications that might be previously unknown to the user. There is already a small task proposed at JPL as a proof of concept to begin testing the concepts of data portability within a Web Services architecture.

The One NASA initiative is a high priority for NASA managers. The idea behind the initiative is to transform the Agency from a highly distributed work environment to a more centralized model. In order for electronic information to flow smoothly from one NASA Center to the next, a consistent information architecture must be implemented as a blueprint for Agency developers through universally accessible mechanisms. The NASA XML Project may be a good venue for these solutions to be developed and tasks relating to content integration will be proposed to them as a possible follow on activity to the delivery of the NASA Taxonomy in spring of 2004.

Technology Drivers for Changing Models of Content Delivery

Recent industry developments indicate that the time may now be right for NASA to consider the implementation of a new content delivery infrastructure. Some of the breakthrough technologies include the adoption of XML and Web Services, the Semantic Web, more implementations of service oriented architectures and the foundation layers of taxonomies and ontologies.

The following sections of the white paper briefly describe some of the leading technical trends in IT that impact content delivery and how these technologies could be implemented in a NASA environment to achieve the goals stated in the 2003 NASA Strategic Plan.

Taxonomy Development

One of the fundamental goals behind taxonomy formation and adoption is to develop a consistent methodology for handling NASA's electronic content. Documents need to be described with a standard classification scheme that follows a predefined hierarchy. This will enable users to see correlations between subject areas. It also allows search engines to retrieve information with more precision and relevancy. Each time an engineer or scientist finds and reuses a piece of content, the return on investment (ROI) of the work to originally produce the material increases 100 percent. 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".

Taxonomies contain descriptors that can be used to mark content chunks. They are composed of discrete branches also known as *facets.* Facets are made up of consistent sets of attributes for labeling content components and can be repeatable. 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. The NASA taxonomy delivery also includes a set of recommended Dublin Core metadata specifications as well as XML schema that will be published with the NASA XML Project's Registry. These products will be freely available to all NASA Centers and Enterprises for use in the building of applications and content repositories.

Due to the fact that it has been designed with a "top down" approach (rather than a "bottom up" approach), the breadth of its classification schema allows the NASA taxonomy to act not only as a means of content identification through the tagging of material, but also as the "big buckets" needed to associate relevant topic sets of information. Hence, for the first time, NASA developers have a means at their disposal for correlating materials from dissimilar repositories by mapping synonymous terms into a generalized framework[5].

The figure below illustrates the use of umbrella terms or "big buckets" to reconcile the numerous information architectures found across the Agency today. Essentially the bottom up approach of individual applications and repositories are now able to be integrated into the top down strategy of a consistent enterprise wide taxonomy.

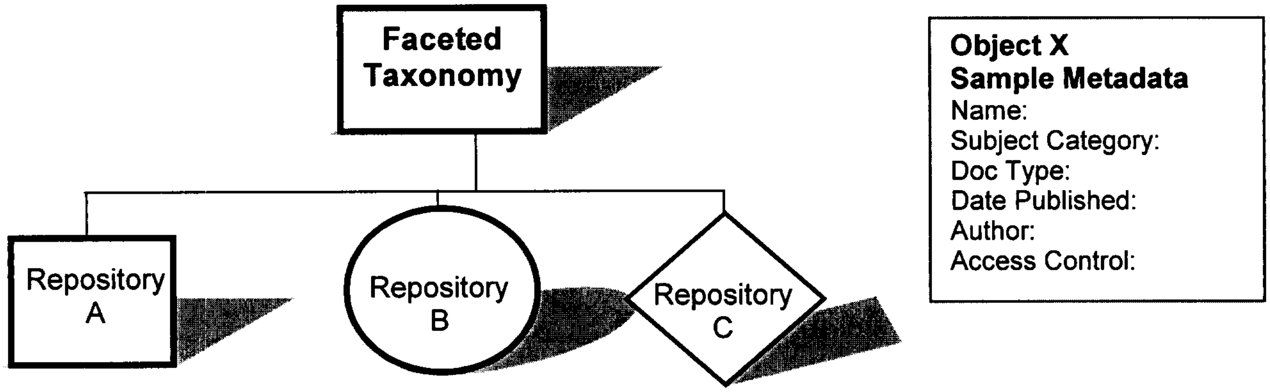
---

[5] *Taxonomy Development With NASA*, Dutra, Busch, https://pub-lib.jpl.nasa.gov/pub-lib/dscgi/ds.py/Get/File-18/NASA_Taxonomy-Dublin_Core_Paper-042203.doc, 4/2003.

Figure 1. Information Architecture From Top to Bottom

Top-Down (NASA Taxonomy)
- taxonomy
- facet /big buckets
- hierarchy
- primary path

Bottom-Up (Individual Repositories)
- sub-site/site
- objects
- metadata
- multiple paths



The Semantic Web

In 1998, Tim Berners-Lee authored the Semantic Web Road Map. In this document he describes an Internet space that has been enabled for use by machines and automated systems. One of his primary assertions is that information available to us through the Web up until now has been *"designed for human consumption, and even if it was derived from a database with well defined meanings (in at least some terms) for its columns, that the structure of the data is not evident to a robot browsing the web."* [6] Through the use of mechanisms such as Resource Description Frameworks (RDF), appropriate metadata and schema frameworks, he proposes the evolution of a Semantic Web that is composed of machine-understandable information.

For information to be found and acted upon by multiple systems, it is necessary to pre-define its scope and meaning. These data definitions are expressed through schema and reside within Uniform Resource Identifiers (URIs) which are easily referenced and found by machines on the Web.

Enterprise Framework Ontologies

Once terms are defined through the use of taxonomies and RDF statements, their relationships to other terms can be specified through the use of ontologies. Ontologies for the Semantic Web are most commonly composed of a taxonomy tailored for the data and a set of inference rules. Taxonomies allow us to define classes of objects and the relations among them. Implemented together, classes, subclasses and relations can be used to express a wide range of information through the use of properties. By allowing subclasses to inherit the properties of their more general parent classes, systems can deduce the proper meaning of derivative terms even if the system does not have a direct link to the original database[7].

Taxonomies can overlap information spaces and allow them to interrelate. This higher meta level of taxonomy formation is expressed through an ontology. An ontology is defined as "an explicit specification

---

[6] Tim Berners-Lee, *Semantic Web Road Map*, 1988, http://www.w3.org/DesignIssues/Semantic.html.
[7] Tim Berners-Lee, *The Semantic Web*, Scientific American, 2001.

of a shared conceptualization."[8] 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.

Inference rules provide the foundation for machines to manipulate terms in ways that are much more meaningful to the human user. In the case of NASA, the Zachman Enterprise Framework[9] supplies an interesting model to implement the concepts of the semantic Web. The design work centers around the analysis of processes used in the core business of NASA: the development of flight missions meant to further human knowledge through scientific experiments and observations. As flight projects refine the processes involved in their development activities, semantic models can be built that describe the business entities and their relationships to each other, including the logistics of needed resources (see Appendix I). The evaluation of processes enables us to specify work flow models that result in robust business products (i.e., a propulsion system appropriate for the mission's science goals, or an instrument designed to capture critical data).

Once the work flow models are defined and documented, much can be done with today's technology to embed content along the way, making it appear at just the right moment in the worker's business routines.

Service Oriented Architectures

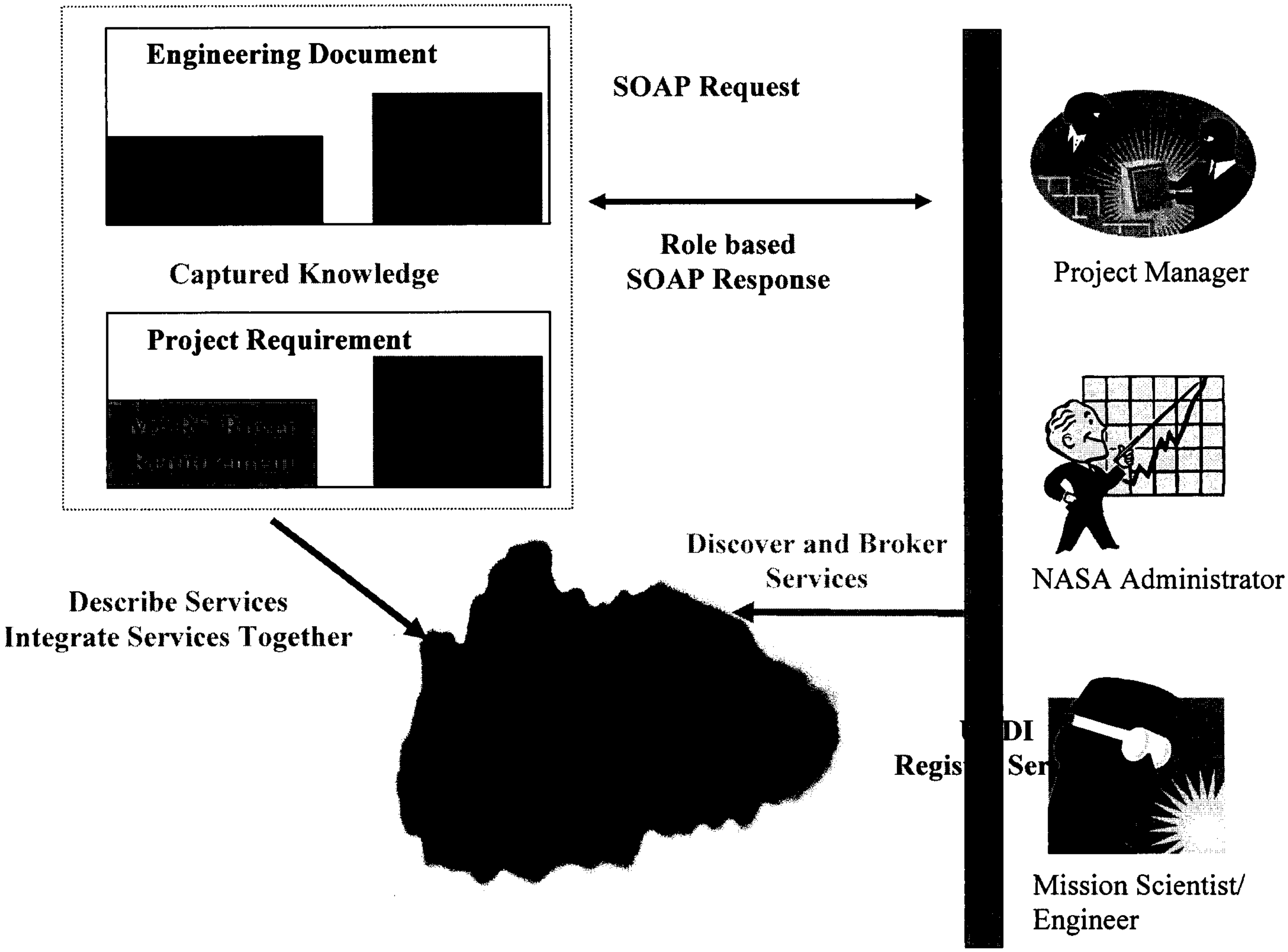
In the past, enterprise applications needing to interact had to be tightly bound with proprietary APIs. These interfaces were built one at a time for specific task enablement and could be easily broken by a change in configuration at either end of the information transaction chain. The new service oriented architectures (SOAs) depicted in Figure 2 are based on Web services. There are three fundamental components to Web services:

- SOAP (simple object access protocol) - the transport layer for XML; it is the means of moving content from one application to another.
- UDDI (universal description, discovery and integration) - this is a kind of central "yellow pages" where a Web application can seek and discover other Web services it may need in order to complete an electronic transaction.
- Web Services Description Language (WSDL) - allows a service to describe how it functions and how another application can invoke its services

Unlike previous interfaces which are usually bolted together with proprietary APIs, the new service oriented architectures are typically loosely coupled. SOAs are self-describing and bind together dynamically at the moment that the components are needed. This provides a more flexible and granular application interface.

[8] *Tom Gruber, Stanford University.*
[9] A Zachman, *A Framework for Information Systems Architecture*, IBM Systems Journal, Vol. 26, No.3, 1987.

Figure 2. Service Oriented Architecture



The Changing Role of Content Management

In the past, content management was thought about in terms of managing a single web site. We wanted to be able to update our sites, control versions of pages and streamline publication through the automation of editorial approvals. Systems commonly include work flow implementation for distributed authoring, multiple display mechanisms, automated posting and archival capabilities.

These days, content management concepts have moved beyond the functionality of single site maintenance to a larger enterprise view of the strategic importance of significant data and better means of delivery, especially for content that is time sensitive or meant to be consumed by teams that are distributed over geographical space. New emphasis is being placed on tagging the content in such a way that it can be delivered efficiently to workers without interrupting their core tasks. The use of XML schema now allows us to add structure to text so that it becomes "portable".

Business Considerations for NASA of New Technologies

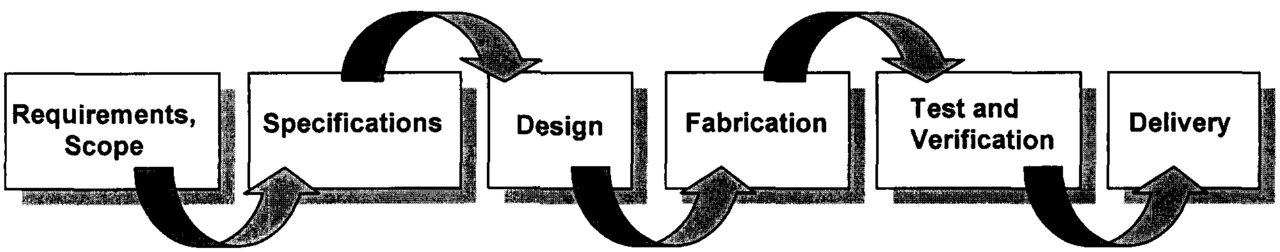
Organizations today are positioning themselves to take advantage of real-time data; by accessing information and tracking its changing nature, they gain a competitive edge in the marketplace. Workers who have access to critical content can see trends earlier and take action in ways that benefit their organization. Although NASA is not a commercial company, it can also use these new technologies to improve the quality and long term relevance of its products. Design cycles for mission development are frequently time-constrained by budget or by a launch window. If the pace of development can be significantly reduced by the implementation of tested design solutions, more new technology can be generated instead of teams spending time on reinventing the wheel.

The suite of technologies described above can have a significant impact on the work environment of NASA by providing information in a coordinated fashion that targets specific content to the individual. In order to accomplish this, an information architecture must exist based on the business processes that NASA considers to be key. When we look closer, we see that there is great variety in NASA business activity. Some NASA projects pioneer technologies and solutions that have never been used before and the nature of these missions may be difficult to characterize without close observation and analysis. Other NASA projects are designed to encourage reusability, such as the Space Shuttle program. However, in both cases, there are patterns and trends that can be determined by stepping back to examine them from a big picture viewpoint.

Using Process to Increase Knowledge Sharing

The critical point for any type of NASA work is that no matter how advanced the technology, the development and execution of mission components rest on processes that tend to repeat themselves over time. For example, in the process of developing engineering products, the life cycle of design rests on the fundamental processes of "requirements definition, specification, design, fabrication, test and delivery". Although the type of development may vary, the development pattern itself is fairly stable and provides a starting point for thinking about mission development methodology in general from an engineer's point of view.

Figure 3. Typical Mission Engineering Life Cycle



The Zachman Enterprise Framework supplies a useful model to expand our understanding of the NASA organization and design an implementation plan of the concepts surrounding the notions of the Semantic Web. The underlying foundation of information architecture centers around the analysis of processes used in the core business of NASA, the development of flight missions meant to further human knowledge through scientific experiments and observations. Semantic models derived from business processes help to describe individual tasks and the relationships between them. [11]

The use of XML allows us to add structure to content previously in an unstructured textual format. By implementing schema with well defined meaning, content can be sliced into finer pieces. This increased granularity of content allows us to move content parts with more flexibility. In addition, once we add metadata that describes business attributes, we are able to begin building an infrastructure that is "content-aware". If we add a layer of inferencing business rules to the data, we are able to build the capability of data streams that are embedded into process work flow.

As flight projects further refine the processes involved in their development activities, semantic models can be built that describe the business entities and their relationships to each other, including the logistics of needed resources. The evaluation of processes enables us to specify work flow models that result in

[11] Zachman Enterprise Framework, See Appendix I.

robust business products (e.g., a propulsion system appropriate for the mission's science goals, or an instrument designed to capture critical data in a particular environment).

Once mission development processes are defined and standardized, we can begin to understand what pieces of information are most valuable to an individual at a particular point in each process. If we have a clear understanding of the process and an architecture that supports the activities associated with each process, then we can begin to find content attributes that will help us make needed data portable and insert the retrieval of such content at the appropriate places in the mission design and development processes. Eventually, we will be at the point where we can integrate content with specific business or engineering applications tailored to a mission or engineering process for real time delivery.

Using Metadata About People to Increase Knowledge Sharing

As we develop useful metadata attributes about content, we are also developing an understanding about the metadata set we need to know about people. Role architecture is in its infancy at NASA, but it is key to creating an environment where the right content can be matched to the right person at just the right point in their work when they need it. Associations and relevancy based on role, work breakdown structures, discipline and access rights will all provide valuable attributes to push appropriate content to various flight teams or communities of practice.

Organizing data to inform mission judgments by providing associated content helps us examine problems in order to mitigate risk and become a proactive, learning enterprise. This allows each mission to build on the missions that have come before. This is the foundation architecture for a true NASA "knowledge base".

Small Tasks to Enable Building Blocks for the Enterprise Architecture

The technologies that enable content integration networks have been briefly described above. When combined with delivery mechanisms that currently exist, we can create new value from tools we already have in place. One example is a proposed task at JPL outlined below to test out some of the concepts contained in this paper.

The proposed JPL task will employ XML Web Services technology as a middle tier layer to create a content infrastructure enabling the dynamic delivery of data appropriate to a user based on his role, profile and specific task through different mechanisms, one of which will be a project portal data channel. Project portals based on the Inside JPL model are now in use by JPL flight teams and, with the development of an institutional information architecture, their true value as aggregators can be leveraged to automatically discover relevant information silo'ed across multiple hosts and diverse applications that might be previously unknown to the user.

The goal is to implement the notion of an infrastructure environment for machines to find and retrieve information relevant to the individual independent of a situation-specific human query, but rather designed to be initiated upon the worker arriving at a certain point in his daily work process. This type of content delivery builds on the ideas of the Semantic Web. If funding can be gained, specific deliverables include architectural components and navigation interfaces designed to perform with high-level usability.

At this time, JPL is developing an Enterprise Metadata Registry Service that is the precursor of an XML Web Services UDDI. In addition, a standardized data dictionary guideline specification[12] has been issued and will soon be adopted. These first pieces give us a central platform to start collecting and mapping various metadata schema from presently isolated repositories.

Conclusions and Summary

Strategic decision making calls for up-to-the-minute information being available at critical junctures. Integrating content delivery into the daily routine of business processes gives workers the additional

---

[12] *Guideline for Building a JPL Standard Data Dictionary*, JPL D-27674, Office of the JPL CIO Information Architecture Project, 1/12/04

perspective of historical data, comparison studies, and contextual relevance. In order for an organization to make the best use of existing knowledge, it is imperative that content be deployed from repositories representing the most significant information resources even if they are architecturally silo'ed. The meaning of a particular piece of data can change radically depending on its context and use.

New technologies are on the horizon, including semantic frameworks defined through consistent controlled vocabularies and loosely coupled service oriented architectures designed to facilitate knowledge transfer. These technologies provide a platform for targeted content delivery to an individual as an integral part of daily work. This is a very different model than the individual stopping the flow of task completion to engage in search and retrieval activities.

The way we deal with information as an enterprise is changing. The role and responsibility of NASA IT organizations are also changing. We are being called upon to do more with less and demonstrate value added implementations of systems that support the knowledge worker of tomorrow. New technologies that can help us realize the full potential of our institutional knowledge are at hand. Our future success as an Agency depends on our understanding the strategic importance of their implementation in order to create a robust knowledge base for NASA.

Appendix I. Zachman Enterprise Framework (Present Day)

	Data - *What*	Function - *How*	Network - *Where*	People - *Who*	Time - *When*	Motivation - *Why*
Scope (contextual) *Planner*	List of Things Important to the Business Entity = Class of Thing	List of Process Business Performs Function = Class of Business Process	List of Locations in which the Business Operates Node = Major Business Location	List of Organizations Important to the Business People = Major Organizations	List of Events Significant to the Business Time = Major Business Events	List of Business Goals/Strategies Ends/Means = Bus. Goals/Critical Success Factors
Enterprise Model (conceptual) *Owner*	e.g. Semantic Model Entity = Business Entity Reln = Business Relationship	e.g. Business Process Model Proc = Business process I/O = Business Resource	e.g. Business Logistics System Node = Business Location Link – Business Linkage	e.g. Work Flow Model People = Org. Unit Work = Work Product	e.g. Master Schedule Time = Business Event Cycle = Business Cycle	e.g. Business Plan End = Business Objective Means = Business Strategy
System Model (logical) *Designer*	e.g. Logical Data Model Ent = Data Entity Reln = Data Relationship	e.g. Application Architecture Proc = App Function I/O = User Views	e.g. Distributed System Architecture Node = I/S Function (processor, storage, etc) Link = Line Characteristics	e.g. Human Interface Architecture People = Role Work = Deliverable	e.g. Processing Structure Time = System Event Cycle = Processing Cycle	e.g. Business Rule Model End = Structural Assertion Means = Action
Technology Model (Physical) *Builder*	e.g. Physical Model Ent = Segment/Table/etc Reln = Pointer/Key/etc.	e.g. System Design Proc. = Computer Funct I/O = Data Elements/Sets	e.g. Technology Architecture Node = Hardware/Software Link = Line Specifications	e.g. Presentation Architecture People = User Work = Screen Format	e.g. Control Structure Time = Execute Cycle = Compon't Cycle	e.g. Rule Design End = Condition Means = Action
Detailed Representations *Sub Contractor*	e.g. Data Definition Entity = Field Reln = Address	e.g. Program Proc = Language Stamt I/O = Control Block	e.g. Network Architecture Node = Adresses Link = Protocols	e.g. Security Architecture People = Identity Work = Job	e.g. Timing Definition Time = Interrupt Cycle = machine Cycle	e.g. Rule Specification End = Sub-condition Means = Step
Functioning Enterprise	**e.g. Data**	**e.g. Function**	**e.g. Network**	***e.g. Organization***	***e.g. Schedule***	***e.g. Strategy***