An Enterprise Architecture Recommendation for Data Integration; A Uniform Information Service Architecture

Andrew Schain[1], Robert Raskin[2], Brian Wilson[2], Rich Keller[3], Walt Truszkowski[4]

[1] NASA Headquarters, Washington, DC, USA
{Andrew.Schain}@nasa.gov
[2] Jet Propulsion Laboratories, Pasadena, California, USA
{Robert.G.Raskin, Brian.D.Wilson}@nasa.gov
[3] NASA Ames Research Center, Moffett Field, California, USA
{Richard.M.Keller}@nasa.gov
[4]Goddard Space Flight Center, Greenbelt, Maryland, USA
{Walt.Truszkowski}@nasa.gov

ABSTRACT:

NASA has an increasingly serious problem managing our data and information resources. Critical information related to our daily functions is becoming more difficult to find and integrate. It is impractical, and nearly impossible, to bring into focus the complete set of applicable knowledge required for understanding and decision-making awareness. The problem is voluminous, diverse, extensive, impacting our entire community, and growing at an unabated rate.

This paper recommends a pragmatic strategy of adopting established global standards and leveraging existing NASA products and best practices to create an agency information service based largely on Semantic Web Technologies. The establishment of this service can be crafted in a uniform approach and will integrate data access across our disparate repositories and databases for customers and machines alike. Moreover, this service architecture will provide our customers the ability to spontaneously search across those repositories and form individualized data collections without expensive and time consuming IT development.

This paper provides an overview of our data problem and suggests specific, technical activities that organize our collective information resources to provide coherence, understanding, and reuse. Complete management control of our entire data collection is a complicated long-term effort, but adoption of this strategy will provide immediate benefits to our entire community and positions us to easily adopt new requirements brought on by our missions, our organizations, and technology advancements.

An Enterprise Architecture Recommendation for Data Integration; A Uniform Information Service Architecture

THE PROBLEM:

(1) Corpus Size & Rate of Growth
NASA has a tremendous amount of data collected over the last 50 years. The exact size and growth rate of our data collection are unknown, as we (employees, partners and customers) are generating new data continuously. Efforts to assess the value of our data collection in either informational or financial terms are difficult, but neither the collection nor its growth rate are likely to diminish significantly in the next 5 years. Today, nearly 13% of NASA's budget is spent supporting information technology.

(2) Variety of Data Sources & Types
Our data and information has great variety in origin, source and type ranging from one-of-a-kind instruments and software, to last-of-its-kind legacy systems. Our collection includes foundational science data and PowerPoint briefings. Our data are stored in man-made appliances and human experiences. Our computer systems and instruments are diverse, spread out across the globe and, in some cases, beyond. Our data consumers are also potential producers, regenerating data through analysis, compilation, edits and emails. Nearly each instance is another piece of data added to our unorganized collection.

(3) Discovery & Relevance
As the quantity and variety of data and information increases, it is increasingly more difficult to find information that you or your organization has collected. Learning of related information outside of your own organization seems impossible, but is required to achieve a more complete and effective understanding of many of our activities. We cannot now anticipate the exact piece of information we will need, but we need to be aware of it nonetheless. In other words, a mechanism is required to present data in relevant context to each unique situation without knowing the data source or potential consumer(s) in advance.

(4) Customer Environment
For NASA, the world is our data and information community. Our customers vary from schoolchildren to university researchers. They encompass nearly all of the disciplines of science, engineering and project management. They speak many different natural languages accented with unique science nomenclatures, technical idioms and the contextual nuances of their own experiences. Even where there is a common language, humans use different vocabularies and meanings, and domain specialists may have difficulties in conveying information to non-specialists. Formalizing semantic mappings across our diverse community of humans and machines is required for our individual contextual understanding and critical for tying information together.

An Enterprise Architecture Recommendation for Data Integration; A Uniform Information Service Architecture

What Does The Future Look Like?

Imagine planning or reviewing procurements and without any prompting your computer discreetly displays a message on your taskbar; "Project Requires Cameras". "There are [5] contracts in place to purchase cameras. Click to view."

Imagine searching for a specialist with a skill set such as an astronomer with a domain specialty in X-ray spectrography and a background in engineering who has successfully managed multi-tiered projects. A graph displays a social network depicting people who match this profile along with connections to the people they worked with. The connections are color-coded to indicate if they are affiliated by project or by organization. Click again and drill down to specific project information or perhaps to see where clusters of spectrographic skills exist. Another view highlights the connectivity steps between you and an element in the graph.

Imagine discovering a potential anomaly on an instrument and mouse-ing over to see systems or subsystems that might be affected, clicking to see the repair or assembly history and mouse-ing again to get contact information for the technicians who performed the work, and science results obtained from the instrument. It is hard to imagine why this service does not exist already today. Why can't we ask through the help of our computers, "who set up this wiring harness?"

Imagine picking from a list of available distributed heterogeneous data and information sources; mission logs, cameras, audio files and instruments; and then assembling everything related to a moment in time and "playing it" back to view all of the inter-relationships within a specific time. Integrate a GPS capability, and imagine replaying the sequence parsed by location or view.

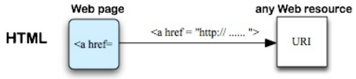
Imagine sorting through our entire collection of NASA photographs to find matches based on a specific coastline. Imagine establishing this association as the west coast of Costa Rica and having it saved and available for others to add latitude and longitude information, hydrospheric data, instrument source data, or geopolitical information; each new source added by subject matter experts provides a continuous enrichment of related knowledge, all while keeping the original photographs stored in situ.

Imagine being able to gather, with a single query, all of the available data bearing on a particular Mars crater, or rock, and the geology of the surrounding region, effortlessly combining in situ measurements from a rover with satellite-based imaging and remote sensing data at several frequencies, all properly aero-registered to a digital elevation model. The layers of information can then be visualized and manipulated in standard

Geographic Information Service (GIS) tools adapted to Mars. Such multi-instrument, data fusion scenarios are critical to deriving maximal science return from the full set of deployed instruments. The many instruments on the three Earth Observation System (EOS) satellites present even more challenging fusion scenarios due to the continuous acquisition of terabytes of hyperspectral climate data. One cannot solve these problems by simply trying to "guess" which combinations of data will be needed to address science questions, and then supporting only those fusions. The data from every instrument (and physics model) must be labeled with metadata that is semantically rich, and the information service must retain and share that rich metadata, so that new, unanticipated combinations of data (and models) can be gathered and fused on demand, at the push of a button.

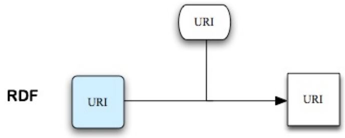
A Brief Primer Of Semantic Web Technology Building Blocks

Many are familiar with a hypertext link that has one end anchored in specific document content and the other in a more descriptive representation that can be found outside of the document.



The basic construct that makes the World Wide Web the most effective information repository ever is the hyperlink, or link. A link consists of three parts: The source document (and source region in that document), the link itself, and the target of the link, that is, what you get to by "following" that link.

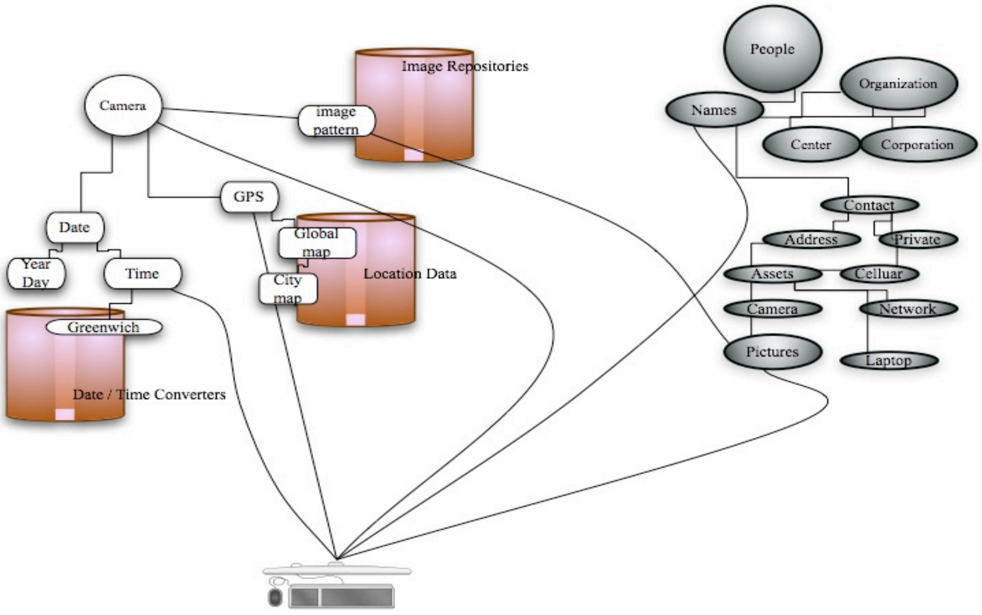
The core Semantic Web construct, as embodied in RDF (the Resource Description Framework), is the "triple". Triples are not just human oriented, featureless links, but *assertions* that consist of a named subject, predicate, and object. Thus, they allow for richer descriptions than the bare hyperlink allows: First, the link itself may be *typed*, that is, there are distinctions between links and links between the same source and target can represent different relationships between them. Second, the link is made *explicit*. Instead of being hidden in the dynamic interaction between pages and the human browser, all parts of the link are made tangible. So, meaningful links can, themselves, be assembled into larger structures. If you want to describe (in natural language) that a particular camera has a specific focal length, you might say: "An Elf 200 – has a – focal length of 5.0". In RDF, the subject – predicate- object (the triple) are each represented using a specific Universal Resource Identifier (URI) – an address. While still anchored in data, the description as well as the entire relationship can now be found outside of the application, out in the world where it can used and controlled in more expressive ways.



Next, we can gather these descriptions and group them together where it makes sense. So an Elf 200 may be member of a group (class) called “Digital Cameras” and “Focal Length” may be a property of both the class “Camera” and subclass “Digital Camera”. These description and assertion mechanisms are ontologies that can be used repeatedly by multiple programs.

We can organize the classes and properties and apply logic, inference, and rules so that ontologies can join to other ontologies when required to satisfy arbitrary customer-driven applications. Not only can you ask, “show me all digital cameras with a focal length of 5.0”, but you can specify “anyone who lives in this zip code gets a 25% discount” or learn that “people who looked at this camera were also interested in carry cases”.

The use of standard ontology languages such as OWL (ontology markup language) or RDF enables us to organize data in small logically contained groups that are machine readable and processable – that talk to each other. The expressiveness in the language far exceeds traditional databases and provides extremely accurate searches with incomplete data sources as well as machine assisted (readable) searches, affiliated queries and step-by-step inferences. In other words, machines will help us to connect the dots.



Example: Query all of the photos taken at KSC between 2:00 am and 4:00 am.
Now, just show photos of Delta 2s taken by Bill.

By applying re-usable machine-processable data organization constructs at the web level instead of hard coding them in trapped databases, consumers can form queries and collections without expensive programming.

Unlike database applications that might have web interfaces, a Web Service can import data directly from distributed applications into your own "web view" with minimal human intervention. Think about entering a zip code to get a customized weather forecast. By entering a zip code the related satellite image and a picture of local radar are automatically associated and displayed, often including local advertisements or news. Or think about the series of tasks that lead to a camera purchase or flight reservation. We can organize the relevant "data-blobs" and services so they have contextual relationships. There are mechanisms that machines and data sources use to connect to each other and form coherent and related step-by-step processes.

One service mechanism on the standards track is a form of OWL (OWL-S) that along with WSDL (Web Services Description Language) enables collections of services and collections of tasks much like collections of data. Services (and even specific devices) can be advertised or discovered on the network as available for your particular needs. Tasks are organized and given real world meaning using ontologies and the associated instructions for action are assigned in WSDL. So an OWL-S description might say what a service is or does and provide a track to associated possible steps. For example, if you are buying a digital camera and enter something fairly opaque to a search engine like "Nikon D70", your output should be a variety of choices including the starting point for purchasing *that* model camera. As you proceed and new choices are offered, the system needs to know the relationship between the steps while keeping track of the specifics (e.g. a particular camera model). That way, if you are momentarily distracted in your purchase to a more powerful lens or additional memory you can always go back to the basic model.

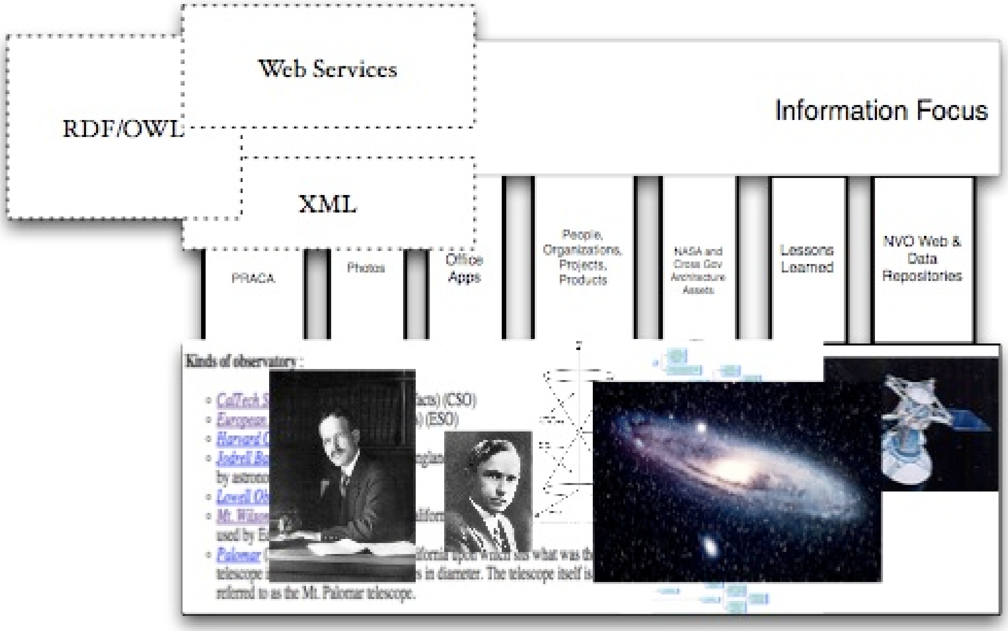
An Information Service Architecture based on the full Semantic Web Technology "stack" will enable our customers to intelligently discover information and relationships, and easily link tasks together. Also, services can be more than just tasks, they can be resources; a service can advertise itself on a network and be made available to customers based on the customer's or application's credentials. For instance, your digital camera could be plugged in to your office computer as a file resource for anyone in your work group. Or scientists could "publish" their working analysis to selected collaborators in the form of an eScience notebook, from which documented analysis algorithms can be reused as remotely-callable services. This service model provides a new solution to the perennial problem of software reuse. Algorithms can be published as services for discovery and reuse, instantly bypassing the usual problems of choice of implementation language and portability of code.

The growing number of callable Web Services in the business and scientific communities has led to the rapid development of numerous workflow engines to "choreograph" or connect together the services to accomplish new tasks. For example, Business Process Execution Language (BPEL) allows customers to tentatively book an airline flight, rental car, hotel and massage, and then commit and pay, or cancel the entire transaction if any service is unsuitable. The innovation is that a person can assemble custom workflows combining services from different companies in unanticipated ways.

Web Service choreography is maturing rapidly in both the business and science domains. Workflow systems range from informal systems that organize work-day activities, often with part of the system still on paper or in humans' minds; to more structured but human-oriented workflow to edit, review, and approve the content that will be published on a web site; to formal production systems that automatically process terabytes of "downlinked" satellite data that produce higher-level products. Sophisticated data processing systems, distributed across the Grid, can be assembled using a variety of workflow engines for choreography, transparently locating data replicas online, and submitting jobs on remote supercomputers. In all cases, there is a happy synergy between semantics and structured workflow: the workflow benefits from semantic metadata and the potential for logical inference; conversely, a structured workflow system provides an opportunity to capture, preserve, and infer additional semantic annotations (metadata). The more structured the task, the easier it is for a computer to capture human intentions. Thus, it is vital that workflow systems be semantically enabled, meaning that they:

- Use semantic metadata to understand (infer) what they are doing and potentially improve the data flow;
- Preserve semantics by saving links to the semantics of (metadata describing) the input datasets, related datasets, and the data transformations (algorithms) used to generate downstream products;
- Generate new metadata by allowing the user to incrementally (or automatically) add semantic annotations to the generated data products; and
- Infer new semantic metadata by understanding and applying logic to the semantics of the data and the transformations performed.

The future for an Information Service built on Semantic Web Technologies—services, workflow, knowledge building—is very bright. New algorithms can be published, discovered, and reused as services; custom workflows assembled using visual programming, and then exchanged and reused; and domain ontologies authored, again using visual editors, and then accessed and queried as services. Each incremental improvement, automatically discovered and used, strengthens the overall system.


Web Services
RDF/OWL
Information Focus
XML
PRACA
Photos
Office Apps
People, Organizations, Projects, Products
NASA and Cross Gov Architecture Assets
Lessons Learned
NVO Web & Data Repositories
Kinds of observatory :

APPROACH

Implementing a plan to manage our data and information is a large undertaking. The current problem is complex, multidimensional, and took years to develop – so it won't be solved overnight or with a single simple answer. Yet there are small, surprisingly simple, well-defined steps we can take right now that will generate immediate benefits across the agency and create an ever widening "network effect". ***We propose a strategy that enables our data and information customers to drive incremental metadata organization based on their needs and, through careful stewardship, each valid construct can be left for others to reuse and repurpose.*** Over time, this approach will solve the data problem and position us to easily adapt to changing data requirements and devices.

1. Establish Project Leadership & Organize.
2. Establish Core Infrastructure and Processes.
3. Establish Attractor Services.

An Enterprise Architecture Recommendation for Data Integration; A Uniform Information Service Architecture

Establish Project Leadership & Organize

Data ownership (stewardship) is everyone's responsibility, but currently no one is responsible for building an integrated data capability across our diverse systems into a single information service. Shrinking budgets and competition between offices makes leveraging work and resources even more difficult without leadership. Leadership must provide assurances that the next steps we take are technically extensible, economically sensible, and culturally practical. Management must be there to provide the guidance needed to encourage beneficial activities and discourage activities that represent the wave of the past. To beat back the data mountain dilemma, a small group of people should be assigned full time ownership of this problem. In NASA-speak, the problem needs to be *projectized.*

Adoption of an Information Service Architecture must be clearly articulated as part of NASA's future in order to encourage the participation of our best and brightest people. Leadership should establish an integrated project plan articulating preferred approaches, measurable milestones, and definable goals. They should also provide a venue for individuals across NASA who have established expertise using Semantic Web Technologies (which for this discussion include Web Service and Task Computing technologies) to participate without competing against each other. Established and well-demonstrated expertise at ARC, JPL, and GSFC should be leveraged while enabling undiscovered pockets of expertise to join in.

Establishment of Core Infrastructure and Processes

Design principles are intended to set the parameters for acceptable development and deployment and to offer insight into the rationale and context of decisions to a wider audience. A set of design principles should be created, published, and maintained for the benefit of agency knowledge workers and our partners.
Candidate preliminary design principles include:

- Data organization constructs (e.g. taxonomies, ontologies, XML schemas) must be reusable and available for computer systems/services;
- Web services must be made available for reuse (and strategies need to be developed to identify service types, applications, and rules governing their availability);
- Yield to the greater concept – even if your focus is more narrow; and
- Under-restrict and quantify the ontology, and then decompose it and keep it small.
- Keep it simple to maximize agility and re-use – semantic web technology is still the web we should leverage our existing web infrastructure.

Similarly, a set of strategic principles should be established. They are intended to provide guidance and rationale in the development of management processes. Candidate strategic principles include:

- Develop a strategy about data curator functions, providing assurance, change control, etc.;
- Keep data and contextual validity close to the data owners and subject matter experts who care about it;
- Accept that some data constructs or services may not be fully mature at the outset but can be driven by subsequent customer use and applied benefit;
- Protect individual privacies as disparate systems become available to wider use;
- Establish a presence on the W3C Semantic Web Best Practices Group and other standards bodies;
- Make existing XML, RDF, OWL, and taxonomies available as a library service, enabling authorized reuse from authoritative sources;
- Publish "go-to" designs so that application developers can model their systems against a standard, leveraging the work that has gone before, and enabling fast track extension and integration to other systems;
- Understand, document and manage to the measurable success criteria for the initial, intermediate and longer terms of the effort.

Early on, the establishment of a few infrastructure components and a handful of governing processes are needed to provide enough foundational stability to foster growth and build momentum. Candidate infrastructure services include:

<u>Data Representation Libraries</u>
XML schemas, RDF, Ontologies, Thesauri, and Taxonomies, will need to be valid, trustworthy, and available for reuse. We will need to establish manageable, semantically rich official libraries for unique Knowledge Representations (e.g., payload processing constructs, human relation constructs, vehicle and instrument constructs) and adopting more universal KRs constructed outside of NASA but certified for our use (e.g., astronomy and celestial mechanics constructs, biosphere, atmosphere, hydrosphere constructs, telemetry and navigation constructs, facilities, computers and other capital investment constructs). This architecture should enable ownership/authorship and responsibility for domain experts to give others confidence and trust through provenance and (more importantly) through successful results. A process for conversion or translation of traditional schemas or corpora will need to be formalized so our repositories of production-worthy ontologies can grow easily.

Service Advertisement Repositories

Repositories must be established for individuals to publish available services that can communicate with other existing services. The goal is to enable our computers to know when a new service has come online, understand what it does, employ its functions as part of generalized tasks, and specify under what conditions the service can be used and trusted. Testing of UDDI (Universal Description, Discovery and Integration) as well as UPnP (Universal Plug and Play) should be undertaken with careful consideration to browsing, discovery, and trust capabilities.

Metadata Collection and KR Construction

Tools that either harvest existing metadata or provide computer assistance in asserting new metadata and populating ontologies should be assessed and some preliminary findings tested against candidate systems. Natural Language Processors that assist in determining likely metadata elements, as well as simple mechanisms for customers to add semantic annotations, should be evaluated in parallel. Similarly, XML schema and RDF constructs should be used to evaluate mechanisms (such as conversion to OWL) for more expressivity.

Maintain the Semantic “Stack”

While applied development and assessment needs to continue regarding choices in planners, reasoners, parsers, conversion tools, natural language processors and library services, we can identify the basic building blocks that applications should adopt. The entry level to integrate into the NASA Information Service should be set at well-formed XML KRs at a minimum. RDF and OWL are preferred and will most likely remain our standard and target development area for the next five years or more. There are permutations of OWL including OWL-DL (Description Logic), OWL-Full, and OWL-Lite. Choices among these versions should be driven by application requirements including performance with reasoners. For Web Services, standardizing on WSDL and OWL-S for the next three years is appropriate. Integrating with UDDI, UPnP, and .Net applications should be explored initially to determine scalability, extensions to other services and policy-based restrictions.

Participation in Standards Groups

Most of the enabling technologies and components discussed above are mature enough for adoption while others are maturing quickly. We should sufficiently staff this project sufficiently to participate in standards and industry groups, to assess applicability of complimentary technologies, and influence the development process to assure that our future requirements are met. Research and industry partnerships should be pursued and maintained for NASA’s (and the public’s) benefit. For example, we should promote building extensions to OWL to perform mathematical reasoning and mathematical representation. We should promote standard data-model APIs including SPARQL and DIG. And we should facilitate “bridging middleware” that enables access to existing information services. These would include:

- o Establishment of SVN/Annotea-like capabilities for KRs;
- o Establishment of tools that will generate RDF from office-type applications;
- o Establishment of tools that will generate ontologies from database schemas.

Establish Attractor Services

The *network effect* describes how a service becomes more valuable as more and more people adopt it. As more services and capabilities get incorporated, it motivates more individuals and more services to participate. The more services we tie together, the greater the utility. The greater the utility, the more services get incorporated.

We cannot achieve an interconnected information service all at once. However, carefully selected projects that are useful and have broad utility will create *attraction* for other projects to connect into the information service. By focusing on less than a half dozen projects, we can quickly provide our customers with benefits now and establish sufficient momentum for a long lasting network effect. The selection of these attractor projects should be based on opportunity, leverage of existing skills, and customer appeal. For maximum impact, candidate attractors should be deployed from several communities within a short period of time. Ideally, we should select both applications that serve science and engineering communities and ones that serve institutional and populist needs.

CANDIDATE ATTRACTOR SERVICES

The following suggested list of services are opportunity driven and can be integrated with each other in a uniform service to deliver on the promise of information management. Some of these attractor services are already part of planned or funded projects and would require only additional advocacy for an integrated approach. Others are not yet funded. This list helps to illustrate the intent, direction and benefits of an organic approach.

1. Linking People, Organizations, Projects, and Skills.
2. Adding metadata search and inference in image inventories.
3. Federal Enterprise Architecture and Capital Investments.
4. Semantically-enriched Document Management.
5. Integrating Science Knowledge.
6. Semantically-Enabled Workflows.

An Enterprise Architecture Recommendation for Data Integration; A Uniform Information Service Architecture

(1) Linking People, Organizations, Projects, and Skills

Finding people either by expertise, organization, skill, or project participation, and drilling down to get details regarding contact information, publications, project pages, and the relationship to organizational structures would be extremely popular for our entire community. We are always trying to find what project or product has already been completed or is in process, and there is no easy way to do this today. This proposed service enables customers to find information about projects (e.g., initiative summary, sponsoring office, affiliated program, participating individuals) and to find out about other work those offices or individuals have produced or published in the past. Formalized inferences can be made that an individual may have certain skills because their education, publications, and participation imply those skills.

This service already has a proof-of-concept constructed in support of the NASA Engineering Network. It is based on well known Internet ontologies (FOAF and DOAP) but modified to exclude bnodes, and classes not relevant to NASA. Authoritative data sources like directories (NISE) and a Competency Management System will be used in the development of a prototype. We will help customers to easily search for any combinations of skills, experience and background by using visualization tools like Mspace and Activespace at the web layer. If the prototype is accepted, POPS will contain all of our ~80,000 civil servant and contractor employees and would be available for applications and internal customers alike and portions accessible to external customers.

(2) Metadata Search and Inference in Image Inventories

NASA has hundreds of thousands of images, stored in multiple formats, at multiple levels of trust, at multiple levels of resolution, with associated descriptive information at multiple levels of detail and formality, and in multiple locations across the country. NASA also generates thousands of images on an ongoing basis that are collected and cataloged often in accordance with the needs of the image creator’s specific disciplines (e.g., principle investigators, mission specialists, public affairs, etc.). Some images are simply named and stored with the defaults of a specific application; providing integration of these data sets will have a very big impact. There are many image annotation tools NASA can adopt now that will enable both providers and consumers to annotate regions of an image using concepts defined in ontologies. There are additional opportunities to automatically import or generate metadata about the content of the images. A proof of concept has already been developed that enables someone to retrieve a photograph of say, a shuttle crew – drag your mouse over a particular region depicting an astronaut and harvest information about that individual (e.g., place of birth, education, other missions, background work, etc.). This model can be extended and integrated to other (nonstandard photographic) images such as CAD drawings, design diagrams, engineering drawings, and so on. Associated budgets or investigative science results could be tied to specific regions of those images.

All of our customers are attracted to NASA images of planets, people, models and machines. Integrating our photographs with data attributes provides a high value service that will encourage participation and integration from earth and space science mission contributors, data owners of engineering models, facilities blue prints, graphic designs, design diagrams, as well as other image content providers.

The potential for large network effects in this area has already been demonstrated by the growth of social image annotation (e.g., flickr) and social bookmarking (e.g., del.icio.us) on web sites. (These sites are fairly unrestricted, but the same principles can be applied to large or even small workgroups.) Once mechanisms are in place to enable users to easily annotate images and data, and to share and query those annotations, humans will naturally do the work, and benefits should grow exponentially. The social aspect of "Annotate & Share" creates a virtuous circle that drives adoption. By providing access to common vocabularies (standard ontologies) developed by domain experts within annotation editors, , the semantics of user-provided tags can be made more uniform. As visual editors for concept maps and ontologies mature, the "social tagging" phenomenon will inevitably progress to "social ontology building" by groups. As with everything on the Web, the most useful (and greatest mass of) information will be created by users reading and authoring, viewing and annotating, adding unexpected semantic links, rather than by centralized institutions.

(3) Federal Enterprise Architecture and Capital Investments

Currently NASA's Enterprise Architecture is being compiled in a database to track (at least initially) our "as-is" IT infrastructure assets. By asserting key metadata concepts against this data repository and organizing them based on already existing NASA thesauri and taxonomies, we can build a reusable Agency KR. By associating networks and other computer assets with locations, we could map capabilities and services. If this service were developed, affiliated information regarding our IT assets could be integrated in. Under the correct constraints, assets such as displays, projectors, printers, cameras, or servers could advertise themselves as available services, possibly linking our conference rooms together. Locations of instruments and equipment that comprise services for testing could have customer schedules and funding sources tied to them. Associated capital investment and depreciation calculations could be added. Schedules of potential re-investment or maintenance costs could be forecasted by leveraging point-of-sale type applications. A query capability to see not only what is available but when and where investments may be needed could evolve.

Interoperability and compliance with other Federal Government initiatives such as the Data Reference Model (DRM) will require NASA to organize and share our information across Federal Agencies. At a minimum, since the data exchanges are currently based on XML schemas, NASA will be able to deliver XML schemas to federal agencies while using the more expressive and extensible view of the same construct for more powerful internal purposes.

(4) Semantically-enriched Document Management

The Agency utilizes numerous web-based document management systems to store its electronic work products, including Windchill, Docushare, NX, VRC, Livelink, PBMA, Postdoc, and others. These systems store electronic files in hierarchically-structured folders, but understand nothing about the nature of the information stored in those files or the relationships among files. For example, aside from a document's title or its presence in a specifically-titled folder, there is no way for these systems to distinguish between a scientific report and a budget report, or to determine that the budget report is associated with the same project that generated the scientific report. This makes searching for information difficult because the search must rely solely on text – whether in a document title, folder title, or document content. There is no way to search based on the type of information being sought or the interrelationships with other information. The whole notion of information semantics is missing from these systems.

SemanticOrganizer (encompassing InvestigationOrganizer and ScienceOrganizer) is a semantics-based repository system developed with NASA funding that applies semantic web technologies to the problem of managing information for a variety of different types of NASA projects, including scientific research, accident investigations, and engineering design projects, among others. This system has been successfully deployed since 2001, has over 500 registered users and has been used by over 40 different project teams including the Columbia Accident Investigation Board. Although modest in scope due to its genesis as a research project, SemanticOrganizer illustrates the potential for semantic web applications within NASA. A next step would be to demonstrate the ability to wrap conventional repositories in use at the Agency with a semantic information overlay, working toward wider-scale deployment of the benefits of semantic repositories. In fact, progress has been made in this direction with the development of NX-IO: a joint project between ARC and Xerox to deploy InvestigationOrganizer on top of NX core services. This system is currently in alpha release and under review by Xerox for further development and ultimate product release to customers such as the National Transportation Safety Board. Since many NASA customers are looking to upgrade past Docushare v.3 and will likely move to NX anyway, there is potential for leveraging the NX-IO effort and incorporating it as part of the standard NX product, thus creating an infusion path for semantic technologies into commonly-used NASA tools. A similar opportunity exists for layering semantic meaning on top of Windchill and other components of the ICE suite; ultimately benefiting ESMD customers by providing sophisticated search, reasoning, and integration services not available within any of the ICE components themselves.

5) Integrating Science Knowledge

Suppose a researcher wants to examine how the El Nino of 1997-98 impacted public health. Where does one begin to locate, access, and integrate data from multiple disciplines? A search for "El Nino" on Google or through the Global Change Master Directory (GCMD) would miss many sources of oceanographic data because the search engine does not understand that El Nino is a phenomenon in the Tropical Pacific affecting temperature, rainfall, coastal fisheries, etc. Similarly, technical public health parameters may be unfamiliar to a physical scientist who uses colloquial terminology. An intelligent search tool would consult ontologies to find how a concept can be alternatively represented, then search on the expanded term list. Humans could browse the concept space as an electronic encyclopedia. Moving the mouse over a concept could identify experts in that discipline. Or pressing another button can download data directly onto a map that overlays multiple products, and displays animations over a desired time sequence. The Earth Observation System (EOS) generates terabytes of data per day. The semantic content of this data is often lost because we are not set up to visualize such large time-dependent archives.

Solutions to many pieces of this problem already exist, but they need to be integrated together. Ontology-enhanced "smart" search services broaden the searched space by generating synonyms and related keywords for free text search and inferring related concept tags for structured search of XML metadata repositories. Modern data repositories offer services to query by data type, time, and planetary location, often in the form of structured, XML-based Web services. The services paradigm, coupled with machine-processable semantics, provides the "smart glue" to tie together free text search, XML/RDF search, time & location search, and other services into an integrated Information Service that supports multi-paradigm queries.

Representation of science knowledge in a machine-readable form already exists as part of the Semantic Web for Earth and Environmental Terminology (SWEET), developed at JPL. SWEET includes an integrated set of ontologies describing Earth science and related data concepts and an associated search tool that does not require exact term matches. The ontology content includes all of the knowledge contained in NASA's Global Change Master Directory (GCMD). Future plans are to expand SWEET to include space and planetary science concepts. The ultimate objective is to enable seamless analysis using ontology-aided tools and services.

(6) Semantically-Enabled Workflows

There is a strong synergy between semantics and structured workflow: the workflow benefits from semantic metadata and the potential for logical inference; conversely, a structured workflow system provides an opportunity to capture, preserve, and infer additional semantic annotations (metadata). The more structured the task, the easier it is for a computer to capture human intentions. Thus, it is vital that workflow systems be semantically enabled, meaning that they:

- Use semantic metadata to understand (infer) what they are doing and potentially improve the data flow;
- Preserve semantics by saving links to the semantics of (metadata describing) the input datasets, related datasets, and the data transformations (algorithms) used to generate downstream products;
- Generate new metadata by allowing the user to incrementally (or automatically) add semantic annotations to the generated data products; and
- Infer new semantic metadata by understanding and applying logic to the semantics of the data and the transformations performed.

Uses of semantics include service description & discovery and interface mediation. By semantically describing services using OWL-S, one can query for, discover, and reuse services that fill missing steps in a desired processing stream. Candidate matches must not only have the correct number and types of inputs and outputs, but must provide a transformation or service with the desired functionality (data subsetting, regridding, mining, fusion, etc.). Ontology-enhanced search is vital here. If service interfaces differ slightly in form or input/output types, semantics can also help to automatically mediate and adapt one interface to the other. Eventually, we will expect workflow systems to automatically discover new services and operators (algorithms) that help us do our jobs better. Which regridding operator best interpolates one climate model grid onto another, properly accounting for the inherent variability of atmospheric temperature and water vapor? Perhaps the one published as a service, with an accompanying paper, by a well-known climate modeler.

Preserving semantics is a matter of saving references to the input data (using permanent object IDs), and saving existing links between the data and semantic metadata (using permanent URIs). Complete provenance for all generated products can be maintained by saving the list of input datasets, auxiliary control data, names & versions of all the operators/services in the dataflow, execution log, who requested the execution, and other traceability metadata. Such provenance metadata is "semantic" if it is richly linked into other metadata repositories. The links can be explicit or discoverable later via logical inference. Thus, "who requested" leads one to that individual's skills and publications, "operator versions" to algorithm documents and the skills of the programmer, "service descriptions" to compatible services, "dataset names" to their own traceability chain, etc. The engine should also offer the user an opportunity to label all of the generated products with additional semantic annotations. Every reminder to Annotate & Share yields benefits.

Automatically inferring semantic metadata is a more difficult task. The SciFlo engine will add semantic annotations (labeled as tentative) to its generated products by applying logical inference to its classification of datasets & operators. If a geographic co-registration operator brings two instrument datasets together, merged, and differenced for comparison plots and statistics, then the output is labeled as a "merged, co-registered, cross-validation dataset involving the two input datasets and their retrieved variables". As ontologies mature and inferences grow more trusted, we will soon expect every workflow system to generate "candidate" semantic annotations for our approval, and be disappointed with any system that remains mute.

A distributed network of SciFlo execution nodes will be deployed this year at several universities and several of NASA's Earth Science centers, beginning with the Distributed Active Archive Centers (DAACs) at JPL, Langley, and Goddard. SciFlo is not intended to replace or compete with the large production systems that ingest raw instrument measurements from the three EOS satellites and produce calibrated data (Level 1), retrieved physical variables (L2), aggregate variable grids (L3), and higher-level products (L4). Its purpose is to enable researchers to inject custom data selection, mining, and fusion operators and services into the DAACs and thereby efficiently generate *custom, multi-instrument* L2, L3, and L4 products for *large-scale* science investigations. Each researcher can have his own SciFlo node to participate in the web choreography, and to serve generated (fused) products to the community. The SciFlo network, and other workflow efforts already underway, should serve as a fertile environment for testing and enhancing the semantic capabilities of workflow engines.

The combination of workflow, provenance, and rich semantics will enable entirely new kinds of information discovery. Imagine a scientist who notices one day that a terrestrial weather model is behaving "strangely" in the tropical Pacific. She may have noticed this by conducting a comparison, over years of data, of a climate model to daily weather analysis from the data assimilation model. Is the anomaly due to a poorly understood cloud physics process in the tropics, or the recent introduction of a new satellite data type into the weather assimilations, or a long-standing but little known inadequacy in the climate model, or a bug in the latest version of one of the codes? To investigate, one wants to query the Information Service for the properties of all of the data, calibration systems, model assumptions, model algorithms, and code versions that contributed to the result. Alternatively, one may want to turn the problem around and find all of the results that might be invalidated by a "fault" in a data calibration system. Both forward and backward semantic "links" should be discoverable. Using such links, she could trace back to who made the latest code changes, the publications describing the algorithm changes, and the complete validation analysis, or search for similar anomalies. She might also request a re-run of the weather analysis, withholding the suspect satellite data, and use compute resources on the Grid to run the job (exploiting algorithms as services).

Smart workflow systems that choreograph services will benefit NASA in many ways by enabling customers to: publish, discover and reuse versioned algorithms as services; rigorously specify reusable analysis flows, publish flows and exchange them with colleagues; implement new composite services by authoring a workflow; query the provenance of generated products; label products with text comments & semantic annotations; trace the effects of data or processing anomalies; modify & repeat large-scale analyses, etc.

CONCLUDING REMARKS

The seriousness of our data problem is reflected in long hours to resolve simple tasks, long reviews to assure validity, and missed opportunities. The problem is growing at an unabated rate. ***We cannot anticipate when or in what combination an instance of information needs to be associated with another and so we must design, plan and implement an information service architecture built for those circumstances.***

Semantic Web Technology is concerned with preserving the meaning and intentions that humans ascribe to data and with providing the mechanisms to form links by meaning or intention, not just by document names, dates, etc. For NASA, we must make our data and information just "semantically rich" enough so that automated or semi-automated processes can make more efficient and effective use of the knowledge that we have, infer new meanings based on situational context, and help us organize. The resulting increase in derived knowledge and new ways of graphically "seeing" information will motivate humans to semantically Annotate & Share. Automated semantic capabilities will continue to improve, but we should not wait for an era of super-intelligent agents. Now is the time to prepare, and the required tools and expertise are ready today. By employing strategies of both automatic and incremental annotation, we can begin to enrich the customer's experience now and gain control in organizing our corpora.

Acknowledgments

A note of thanks to the people at NASA who read and commented on a much earlier version of this paper when it was just a jumble of words and thoughts but still offered encouragement to press on - especially Jeanne Holm and Scott Glasser. Special thanks to John McManus, whose advice and support have been a positive constant during the creation of this paper. Thank you to Rob Winters for copy edits and observations. And thanks as well to the wonderful MINDSWAPers; researchers at the Maryland Information and Network Dynamics Lab which is part of the University of Maryland Institute for Advanced Computer Studies. Particular thanks to MIINDSWAP's Kendall Clark, and Bijan Parsia whose careful explanations as to what is real and what is really real were enormously valuable as were their reviews of this paper and their natural way of explaining the differences between right and wrong. And of course my sincere thanks to MINDSWAP's Director, Jim Hendler, who always made himself available for much needed coaching and who remains a resource of immense wisdom, perspective, knowledge and vision.