Business Integration Driven by Business Lines:

A perspective on the Data Reference Model as it relates to Cross Agency Challenges.

Standards Based Architecture to Support Federated Data Management.

Concept Level WHITE PAPER

Developed for the Federal Enterprise Architecture
Program Management Office (FEA -PMO), Federal CIO Council,
NASCIO, and Public Cross Government Initiatives

Industry Advisory Council (IAC)
Enterprise Architecture SIG

May 28, 2003

DISCLAIMER

While the Federation of Government Information Processing Councils/Industry Advisory Council (FGIPC/IAC) has made every effort to present accurate and reliable information in this report, FGIPC/IAC does not endorse, approve or certify such information, nor does it guarantee the accuracy, completeness, efficacy, and timeliness or correct sequencing of such information. Use of such information is voluntary, and reliance on it should only be undertaken after an independent review of its accuracy, completeness, efficacy and timeliness. Reference herein to any specific commercial product, process or service by trade name, trademark, service mark, manufacturer or otherwise does not constitute or imply endorsement, recommendation or favoring by FGIPC/IAC.

FGIPC/IAC (including its employees and agents) assumes no responsibility for consequences resulting from the use of the information herein, or from use of the information obtained from any source referenced herein, or in any respect for the content of such information, including (but not limited to) errors or omissions, the accuracy or reasonableness of factual or scientific assumptions, studies or conclusions, the defamatory nature of statements, ownership of copyright or other intellectual property rights, and the violation of property, privacy or personal rights of others. FGIPC/IAC is not responsible for, and expressly disclaims all liability for, damages of any kind arising out of use, reference to or reliance on such information. No guarantees or warranties, including (but not limited to) any express or implied warranties of merchantability or fitness for a particular use or purpose, are made by FGIPC/IAC with respect to such information.

Credits

This paper is the result of work coordinated under the Enterprise Architecture (EA) Shared Interest Group (SIG) of the Industry Advisory Council (IAC)

Authors

Michael Lang, MetaMatrix

John Dodd, Computer Science Corporation (CSC)

Venkatapathi Puvvada, Unisys Corporation

Lawrence Henry, Computer Sciences Corporation (CSC)

Rob Cardwell, MetaMatrix

Other Contributors

Mohammed Dolafi, Unisys Corporation Jun Lee, Unisys Corporation

Karen S. Brown, NSA

Bruce Peat, eProcessSolution

Bob Greeves, DOJ

Jishnu Murkehy, HP

Neil Chernoff, Computer Sciences Corporation (CSC)

Steve Battista, Cairo Corporation

Tiba Parsa, NASD

Chuck Mosher, Sun

George Hogshead, BearingPoint

Eric Sweden, NASCIO

Mike Ryan, State of Minnesota, NASCIO

Table of Contents

Table of Figures

Introduction

Context: The Federal Enterprise Architecture (FEA) being developed by the Office of Management and Budget (OMB) FEA PMO is a comprehensive, business-driven framework for changing the Federal government's business and IT paradigm from agency-centric to Line-of-Business (LOB)-centric. The FEA includes five reference models:

- Business Reference Model (BRM)
- Performance Reference Model (PRM)
- Service Component Reference Model (SRM)
- Information and Data Reference Model (DRM)
- Technology Reference Model (TRM)

This paper addresses some of the key architecture elements required for the Information and Data Reference Model. It links to other associated recommendations involved with Strategic Interoperability and Business Line Implementation.

The paper does not represent a complete approach to Enterprise Data Management and its ties to Enterprise Architecture. It focuses on three topics:

1. How to discover and organize for sharing between governmental and non-governmental entities
2. How to model and represent information
3. How to access and exchange shared information operationally.

Problem: Government organizations are collectively one of the largest data and information providers but are not organized for internal information sharing. There are a range of data modeling and management needs. We have not focused on all the situations but have focused on an approach to information sharing that can cross organizational boundaries. In general most governmental entities think about information in their own specific context. This attitude institutionalizes itself as:

- I need data
- I gather data myself
- I will own the data myself

The end result of this process is enormous data and process duplication, stove piping of data and an inability to interchange data between different organizations. This is true not only between Agencies, but also within Agencies.

Federated Model: The first part of this paper discusses the needs for data modeling and how with federation and modeling along business lines the information and data models can evolve and be examined from a business centric point of view. This is not done from a purely technical perspective (e.g. storing and accessing the data) but, rather from the perspective of the virtual "information communities" that share the common business goals within the lines of business that exist across various government agency boundaries. The process of gathering information into these communities is referred to as the "Federated Data Model." The term "federated" is used since the communities represent a federation of joint interests around a specific set of information.

It is our assertion that these communities are based, in large measure, on the Business Lines described in the FEA Business Reference Model. For example, all agencies engaged in "*Criminal Investigations*" would have an interest in similar information and in the ability to share this information securely with each other as the need arises. This need is immediate and of critical importance to serve the citizens effectively. Each of the Business Lines described by the BRM is supported by associated:

- Performance Requirements (PRM)
- Data/Information (DRM)
- Services-Components (SRM)

As one can see, these are all very much interrelated.

Need for Core Data: In addition to the Business Line communities, it is anticipated that there will be a community at the Federal level to represent the common interests that cut across Business Lines. For example, there needs to be a core set of information about "Citizens" which is shared across business lines. Similarly, all the business lines have a joint interest in common information about business and other organizations. The paper suggests a set of common data elements and an approach to evolve those core elements based on existing government efforts and a focus on the six focus business lines.

Across Boundaries of Government: The paper makes a strong case for opening the information and data modeling process to non-Federal organizations and to use associations such as NASCIO, NACO, and the IAC along with specific industry associations that fit with the business lines to become active collaborators. This open collaborative approach was used very effectively within individual projects such as the Global Justice initiative and such initiatives as the Y2K activities and can create a more interactive and trusting environment that is critical to the

government transformation envisioned and demanded by the "budget issues" at Federal, state, and local levels.

Business Focus: The paper primarily focuses on the business imperatives for federated data models and is not technology centric. It addresses the process and community building first, with technology taking a strong supporting role to leverage unique opportunities that the latest tools provide. This paper also touches up on why these models are needed and how they will lead to the goals on improved information sharing in a rapidly changing world.

Meta Data Models: The intent of data modeling is to understand the "as-is" data state and the needed "to-be" state that support the scenarios of operation defined by the strategic enterprise and operational plan. Described is a common process to allow all portions of each Information Community to have a shared understanding of core information elements.

The data models developed with use of the core data reference models and core business line elements must be able to answer several critical questions:

- What level of data modeling is needed to support Federated Communities?
- How can we build Line of Business focused services without locking them to the underlying physical information and proprietary architectures?
- How do we minimize the impact to on going operations?
- How do we plan for and support Data Transformation and parallel execution of systems?
- How do we insulate the LOB focused services from evolution that occurs within the data focused IT infrastructure?
- How do we assure security and privacy?

The information and data reference models are metadata that consists of two environments:

- Modeling and analysis environment
- Operational data support environment

The metamodels include the artifacts involved in the strategic planning, business analysis, and ongoing active management of the enterprise data management resources. These artifacts provide the context and the traceability as well as the base set of bi-directional transformation paths that go from the registries to the repositories. The context information is comprised of related business goals, user profiles and security attributes that are attached to what we are calling *"Resource*

Containers". Initial starter Meta models and key packaging and management have been included in Appendix A.

Technology Challenges: The paper also defines a number of challenges and issues that need to be addressed. It provides a technology discussion in Appendix A showing how the information and data reference models can support modeling and analysis while being linked to the operational environment. It also discusses the critical problem of evaluating changes across organizational boundaries or domains, and introduces the notion that changes can be mapped to transformation paths and as a result provide for more automated data migration and schema evolution.

Implementation: The last piece of this paper focuses on operational aspects of the data models. *A roadmap to implementation is one of the key elements of this paper and highlights a practical approach taken by the authors.* The following two approaches are discussed:

- Over the short-run, how do we provide "immediate" access to information in the existing physical environment?

- Over the long run, how do we re-architect the information environment to take advantage of the advances occurring in the XML community?

Short Term: In the short-run, there is an existing infrastructure of information that represents a large investment and will consequently be in here for a long time. There is a need to provide a mechanism that allows this information to be defined and shared by the Federated Data Information Communities, regardless of its imperfections in structure and organization.

We advocate the use of open standards, to describe how the information can be organized and accessed using a standard such as Meta Object Facility (MOF) capabilities that were originally defined for CORBA. These capabilities allow the common data models created by the Federated Data Information Communities to be directly mapped against existing physical data structures for use by the community.

This process addresses the real world situation we currently face, but does not represent a long-term, solution. It is a good transitory strategy, but essentially papers over the existing warts in the information architecture and provides a much needed abstraction layer that enables evolution.

Long Term: Over the long run, the XML community has been evolving mechanisms for discovering information, business processes and sharing such information and processes. Examples in the Federal Government include the work being done by the CIOC XML Working Group and in the commercial world on ebXML. We advocate a long-term path for the Information Communities to

utilize XML based services to supplement and possibly eventually displace MOF based services.

We recommend that the Information and Data Reference Model be linked directly with the Business Reference Model, Performance Reference Model and to associated Service Component Reference Models.

Purpose

The purpose of this paper is to present concepts for the Information and Data Reference Model (DRM) based on Federated Data Management using Model Driven Architecture and Integration. It also points out specific issues related to its integration with and support of the Business Reference Model and elaborates specific recommendations and actions that should be taken. The DRM is designed to provide an understanding of the many forms of information needed, what is available within the enterprise as well as the business lines to which the enterprise belongs. The primary focus is at the conceptual and logical level with an aim towards defining the information needs and rationale for those needs from the user's perspective. We provide a glimpse of the "how to" that we believe will be helpful in planning and business case development. The DRM as envisioned will be used in the Discovery and Definition phases of the process described in the related Business Line Architecture and Implementation IAC paper and the related Strategic Interoperability paper regarding the conceptual and business line steps (Reference: www.iaconline.org). The DRM envisions that there is a small CORE of common data elements along with other common data elements that are needed based on the users perspective. The four perspectives are: the citizen perspective, the business perspective, the government partner relationship with related government employees that are working around the same business function and the internal government owners and users that relate to the portion of the Business Reference Model they are playing. The DRM as envisioned is linked to business process models, the related performance and value models from the perspective of each of the stakeholders. It envisions a set of services related to common information connection and access needs. A high level set of modeling approaches that can be used for the variety of information and data types are needed along with a complete guidebook on how to do conceptual and logical modeling.

Scope

Some of the key implementation aspects of the Data Reference Model are only briefly described. A few of the key issues related to the DRM's integration with, and support of, the Business Reference Model is briefly discussed. These issues include the types of services needed to enable the business line: data integration, data quality, data stewardship, and data migration. These issues need to be addressed more completely as the concepts here evolve based on building out the data and information model along business lines. There are other situations for data modeling and management that are both simpler and more complex that are not addressed within this paper.

Benefits

The primary benefit of the approach described in this paper is reducing the cost and increasing the effectiveness of government though business line alignment of its data while minimizing the reduction in cost and risk that this alignment will cause.

This approach used in this paper is a phased and evolutionary one. The federated data model allows the business of the government to continue to function while we lay a logical foundation on top of it so that the government can transition to a true line of business oriented structure. This caution is essential since the government needs to continue running and the costs and the risks associated with shutting down business lines of the government while it is being re-aligned is unacceptable (e.g. Defense and National Security operations or Air Traffic Control). Therefore we suggest in this paper a method of allowing the government to take what they currently have in data products and logically align them to appropriate business lines to leverage reuse. This leveraging and will reduce the cost and increase the effectiveness of government.

While this paper provides simple templates, registries, and repositories keep in mind that it is a much better approach to take more of an integrated information, data, and content perspective. The only effectively way to accomplish this integration is by linking your current Enterprise Data Management with Information and Data Architecture that is based on the business line goals derived from the business and government mission perspective. Without this linking, one will not see the benefits of efficiency and effectiveness that the government needs.

Our recommended approach addresses some specific goals:

- First, understand the current information and data and support analysis for overlap, quality issues, and planning for business driven changes
- Second, support the operational data environment that entails information integration and interoperability with trust and operational recovery
- Third, support the evolution of data and information design

Our approach introduces a viewpoint to consider the overall enterprise and cross-enterprise data management elements.

Our approach will move the information and data to a common core set of data and provide guidance, including a recommended standards based approach to modeling and metadata that encourages collaboration among and between business lines. A combination of sophisticated tools can be used for a deep data modeling and understanding while a set of simple data-information-content templates are recommended to uncover business line agreements and planning on

needs, gaps, and dependencies between members of the business line communities.

Audience

Chief Information Officers, Executives, Business and IT Managers are primary audiences for this paper. Information and Data Analysts will get a high level look at the new concepts, but many of the details are beyond the depth of this paper. Pointers are provided for the Data Management leaders but it is hoped that more open dialog among the government agencies, especially across business lines will ensue.

Federated Data and Information Reference Model Framework

This section discusses the concepts of "Federated" reference models and describes our recommended framework for achieving secure information sharing across business lines and various government agencies. This framework includes, but not limited to

- Business Line Oriented Data Models
- Federated Registries
- Business Line Specific Extensions to Data Models
- Federated Model/Data Ownership
- Model Driven Data Integration

Federated Data Concept: Federated Data Management is a method for managing and accessing information, data and metadata across physical boundaries along business lines in a secure manner. The physical boundaries might be the result of political, system, department, or enterprise separation. Irrespective of the boundary type, federated data management provides a means to homogenize the data access so that these boundaries appear to go away and the data can be shared among the business line partners. Once implemented, an idealized federated environment enables a business line user to perform his/her function, seamlessly accessing, sharing and analyzing information, without regard for physical storage issues. This characteristic is crucial to effective and efficient government performance. The federated data management framework fits with the Business Reference Model and associated data with core government and administrative functions and the information, data, and content used by the business lines.

Federated Data Components: A fully featured Federated Data Management system may include many of the following business and technical components / functions:

- Model and flow of the data across Information value chain along the business lines

- Shared ownership and data conflict resolution agreements
- Data accuracy and integrity authentication and certification
- Location and identification of trusted data sources and usage profiles
- Security and Privacy requirements and risk assessment matrix
- A facility to capture models of diverse physical systems using constructs native to each physical system
- An integrated modeling environment.
- A framework for capturing relationships between components of the models
- A standards based exchange format
- A Federated Data Dictionary that captures all of the entities in the models of the systems
- An abstraction layer that repackages components of the physical systems to better support the Line of Business mission
- A middleware based execution engine that consumes the information content of the models and integrates information from the diverse sources represented in the models

Model Driven Federated Data Architecture: This architecture supports two overarching goals: understanding the structure, entities, and relationships in a diverse collection of information systems, and providing the information necessary to drive information integration engines. We refer to these two capabilities as Modeling for Understanding and Model Driven Integration.

Federated Data Management incorporates modeling tools that import existing schema's to capture the "as-is" model of an information management system. The tools can then support re-engineering or forward engineering of new cross-systems or cross-enterprise views that can integrate and provide information sharing across the business lines. Modeling tools and repositories share their models using the XML Metadata Interchange.

Building new models for cross-agency or business lines by "hand" without automation support is impossibly complex and error prone. These model-based tools are the heart of what is often called model-driven architecture. But one of the keys is that the models are standardized and can be shared and can support model-driven integration. The models are re-useable components that generate tremendous ROI over time.

Modeling standards for metadata and data management from standards organizations such as OMG, W3C, and OASIS should be the basis of the Information and Data Models. Government leadership and involvement in the maturing of these standards is a critical success factor.

Federated Data Model Example: A high level example of how this can be used is described below. Business Lines or emerging organizations such as the Department of Homeland Security can use the federated approach to view and analyze information and data along a series of business lines. In this scenario, we have considered both the modeling and information integration process and combined those into a Concept of Operations as described in detail in Section 3.0. This example federated data process flow begins by gathering the existing "as-is" schemas and loading them into a Meta modeling tool to gain a Model Understanding. A set of modeling conflicts and gaps will be identified along with the needs to meet new business goals. This will all be part of the discovery process. The approach to address conflicts and resolve gaps will be part of an Enterprise/Business Line Data Management plan that will align with other projects and changes. It will include a combination of incremental improvements and other areas that may be significant replacement. The data, information, and content changes will be business driven. One of the first steps is to create an assessment against the "to-be" Vision based on a set of views for information that need to be shared at the Enterprise or Intra-Agency or cross business line architecture level. An Information Value Model will be the created that will focus on high value information, data, and content. The data, information, and content will be grouped into containers and connected to a set of business line communication channels.

Those views create a new set of "federated" information and data models that can be searched, accessed, and "appear" as an integrated information source. There will be a federated set of registries and repositories that store both "as-is" and a set of new model views; including change management and security policy mechanisms built into the containers. The federated approach makes the "containers" of the information and data visible and cataloged in the registries and repositories while the "assets" are owned and updated by the responsible agencies. The process used to evolve and refine the data modeling understanding and for model information integration allows for an evolutionary process. This approach can evolve but the first step of discovery and high-level business line agreement provides a true understanding of the information and data assets from the business point of view. It captures accurately the "as-is" state. The Value of the Information and the needs from the business-mission perspective with implementation assigned to the responsible business line partners. The result can be a high level model such as shown in Figure 1.

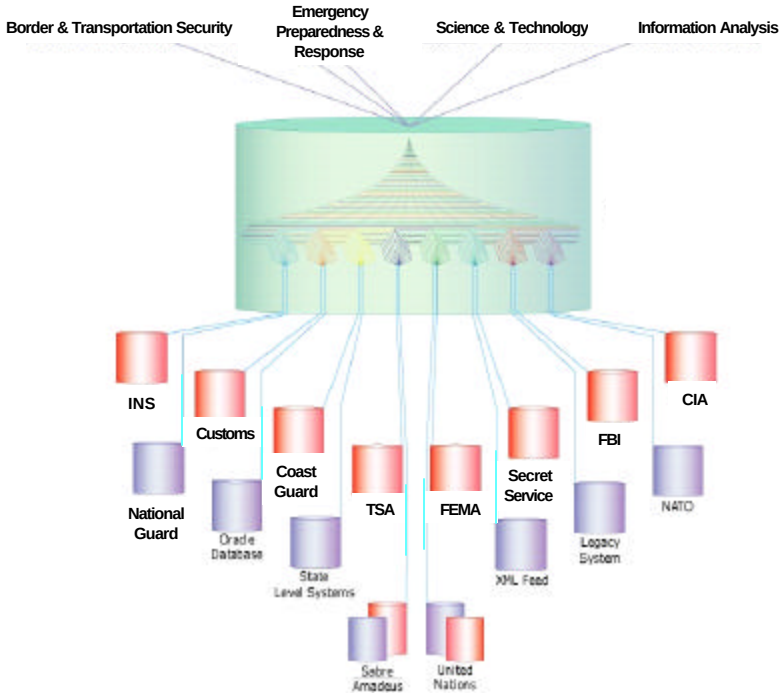


Figure 1: Inter-Agency Information Federation: Critical element of Business Line Architecture

2.1 Segmented Data Zones in a Business Line Context

The breadth and depth of the Federal Government and its many connections to citizens, government partners, and to its business partners introduces great complexity. In Figure 2, we have depicted data models into 3 structured zones in approaching our design of the federated data reference models:

- Citizen Specific Business Line Elements (Citizen Facing Context)
- Government Core Elements (Central Ownership)
- Business Line Core Elements (Business Line Ownership)
- Business Line Partner Extended Elements (Partner Ownership)

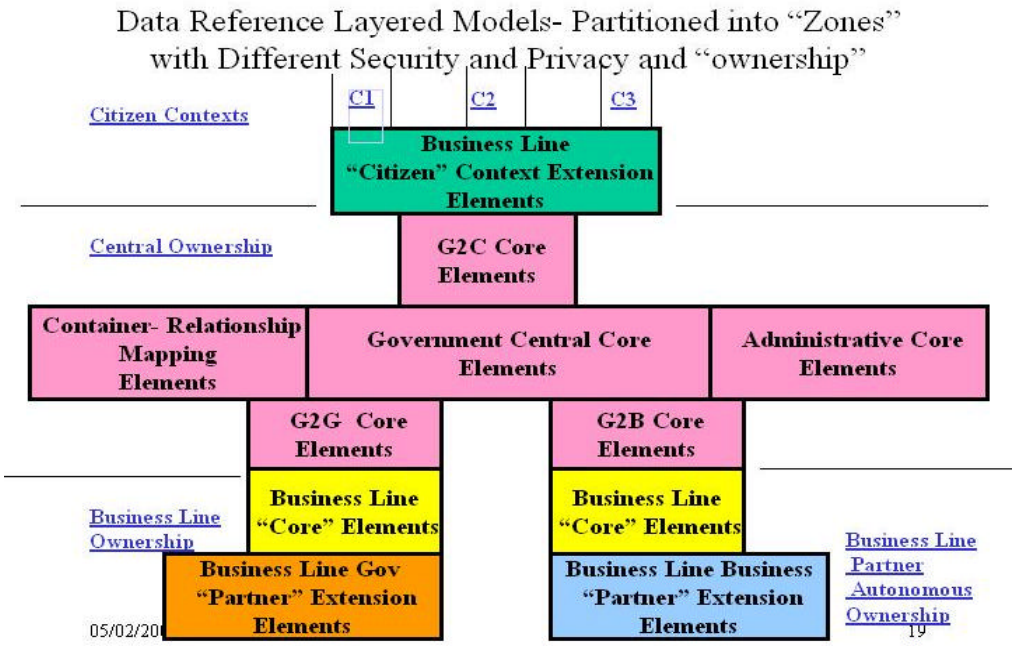


Figure 2: Data Models aligned by ownership zones - Thus Federated Approach

A set of business line oriented data models that are federated within themselves and with a government central "kernel or core" shown in the middle of Figure 2 with three types of data zones for government to citizens, for government to government business lines and government to business lines. By organizing the data and information models along business lines there are a few common core elements across all business lines but most of the data and information is owned and resides along the business lines themselves.

The benefit of this approach is that the data models can be related to the business functions and the business processes and the ownership and zone of trust can be along those business lines. The data and information models can evolve and the changes to the information and data can be defined in business line agreements. Data, information or content entities can be grouped together into containers or collections. We will use the term container to mean any of many different types of physical instances. Containers can be labeled in a consistent manner such as a Core Government Container (CGC XXX- YYY which can be translated into a Universal Resource Identifier) or by the Business Line Container (BLAA- BBB-CCC) with a standardize approach to link the containers together for management purposes along business lines or within larger government information maps.

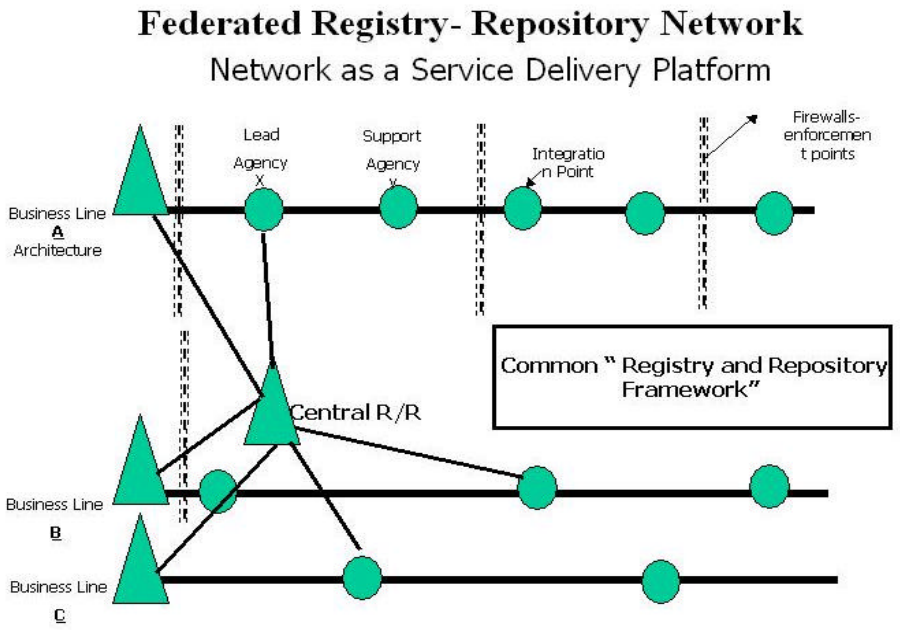
Federated Registries and Repositories by Business Lines

The concept of registries to discover metadata and repositories to access data is very important with in the Federated framework to enable information sharing. Currently, there are a number of registries and repositories projects within the federal government and a number of early adopters in data standards, defining

namespaces, and working with the cross-agency counterparts. The size and scale of the Government information, data, and content can overwhelm the pure central model and the pure decentralization will not provide the information interoperability that is needed. A set of interoperable registries and repositories along each of the federated business lines can provide a service framework for both Modeling for Understanding and Modeling for Integration.

An example of the Federated Registry-Repository Network is shown in Figure 3.

Figure 3: Vision of Federated Registry and Repository Framework



Modeling for Understanding: Building Data Models with Core and Business Line Extensible Elements

One of the key concepts around schema integration and evolution is to have a combination of local views and global views. The federated approach will have local views within the "domain" involved within a business line and one or more global views. Building and evolving the local and global schema must leverage the many excellent projects such as the Global Justice initiatives, the initiatives within DOD, within many of the e-government initiatives such as e-grants, e-vitals, benefits, and the many initiatives that have not as yet surfaced. Common elements for all government business processes or for each type of "delivery mechanism" including government to government, government to citizen, government to business or administrative core elements that primarily improve internal efficiency and effectiveness. Each of the business lines will have its own extensions along with the organizational specific extensions. These possible linking of data, information, and content definitions are shown in Figure 4.

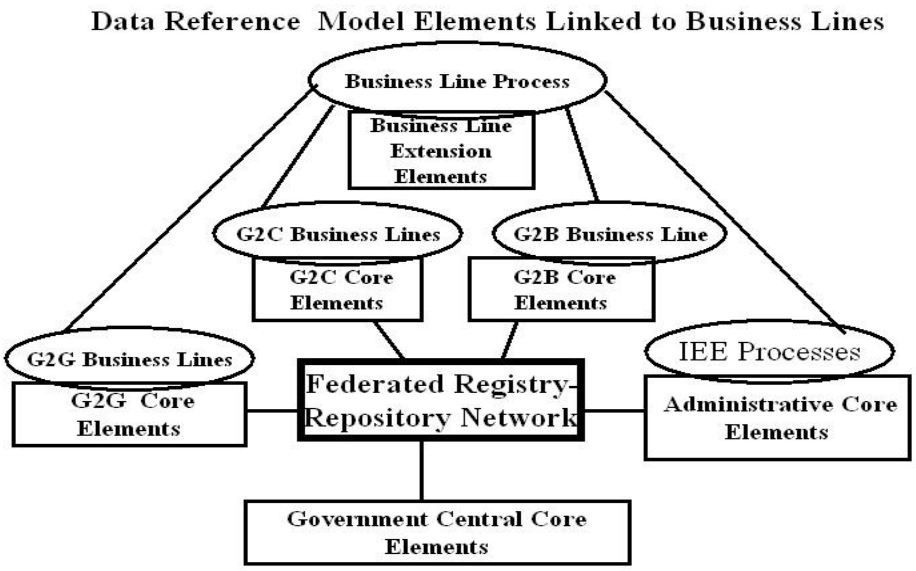


Figure 4: Information, Data and Content that builds upon Common Element Definitions

Federated Data Ownership: Sharing for Success

In a Federated data model, data is owned by the appropriate functional entity along the value chain based on the primary responsibilities agreed with in the business line information flow. As described in Section 2.1, the ownership is segmented in to various zones into core and context specific elements that may be unique and private. This presents an opportunity for common data sharing services and there by eliminate duplicative and disparate data systems. We see this to be an evolutionary process, as this requires a transformation of government culture. We recommend that government establish a working group across each business line to enable this transformation.

Modeling for Integration

There are a number of solutions in use today to achieve some level of data integration: extraction, transformation and loading (ETL) software, data warehousing (DW), enterprise application integration (EAI) and a constant stream of new enterprise information integration (EII) offerings. They all have advantages and drawbacks. The concepts and technologies foran enterprise information integration approach are rapidly converging. The vision reflects the maturing of integration technologies into a more complete solution that synthesizes the "point solution" integration technologies. A model-driven EII based technology that can be the "bridge" between the many reference models and support the transition between the Enterprise Architecture and the Integration will allow the Enterprise Architecture to become a "living EA". It would provide a path between the abstract business functions within the Business Reference Models to applications

that are shown as business process models and services models and connect to the data components that are needed.

Virtual Metadata: What is needed is a method as shown in Figure 5 to link the abstract business needs of the enterprise information consumers to the resources at the bottom the physical data sources. Defining the models and the access paths will require the creation of a new class of "data" entities – virtual metadata (VM). These virtual models can be understood by the end users but must include composition and access management services that allow management and control among the business line partners. These virtual management models and policies become the integration component of Federated Data Management.

Virtual metadata models and their use are a next step in the evolution of data management. We have gone from: the file system to the database and now see the need for a more conceptual abstract model that can be used by the business and mission personnel. Each level of abstraction hides the complexity but at the same times provides the links to next level of information. The series of models and links can provide an information integration path. The definition of data models and the design steps can be linked to the operational environment with model driven middleware.

Virtual metadata is not bound to a particular physical instance of a data management system, but runs in "middleware" and can be used to represent any sort of data system structure. The Figure 5 shows the result of the creation of virtual metadata that can be used for model-driven information integration and allows for increased adaptation to constantly changing needs.

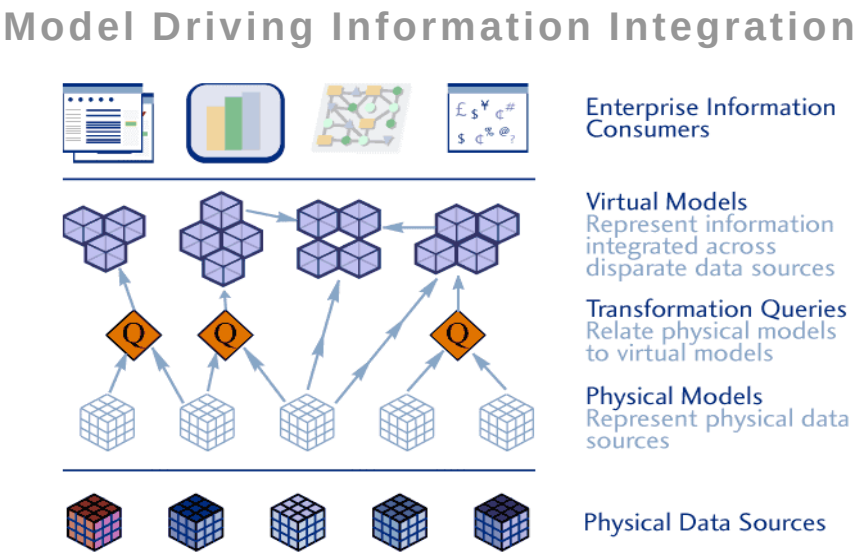


Figure 5: Virtual Meta Data Concept

The process used to create Figure 5 uses a hybrid process approach that includes the understanding of the elements at the bottom viewing them as a diverse collection of physical data management systems. To understand the "as-is" data architecture a set of platform independent models are created for those systems. A set of upward transformation paths can be created which involves the transformation and mappings applied to create virtual models whose entities may or may not exist in any physical system. The creation of these virtual models is also driven by a set of business needs and possible user scenarios. These tell the "Virtual Business Data Model" what to focus on. The resulting process creates virtual models may alternatively be called virtual databases, business objects, views, or enterprise data models. The critical point is that the definition of these virtual entities is created in a modeling tool. The definition of the metamodel and associated set of virtual metadata can be done consistently based on the MOF standard and the resulting standards exchanged with XML Metadata Interchange enabled tools.

Data Integration Principles: The vision of a data modeling environment that supports information integration would thus include models that are:

- all MOF based,
- include relationships models among the "containers or collections"
- provide multi-level models for "as-is" and "to-be" states captured in the modeling tool,

- and includes models needed for the integration process and the change management process.

Managing change as the physical sources change, migrate, or are replaced is then a modeling function and not a custom coding function as the virtual metadata has completely abstracted away the physical systems. The application code that the LOB created to present its data (perhaps a COTS product) executes against the virtual entity so that change in the physical system is accounted for by remodeling that system to its appropriate virtual entity in such a way that the existing application code continues to run unaffected by the change. The business line owners and business line agreements and plans must include a business line data management life cycle and related set of governance and management practices.

Datatype Metamodel: An important byproduct of this modeling process is the capture and creation of a Line of Business Data Dictionary that contains all of the data objects across all of the systems that have been modeled by each line of business. This feature is supported by a specific metamodel, a Datatype metamodel. This enables searching across all of the systems for specific or redundant data types, as well as enforcement of particular data types application use by mapping these "global" data types to the virtual metadata models. These LOB data dictionaries can be rolled up by a higher authority, OMB for example, into a Federal Government Data Dictionary and be integrated with the Registry-Repository effort lead by NIST. A likely scenario is that there is not one global data dictionary but a federated set of federated data registries and repositories as shown in Figure 6.

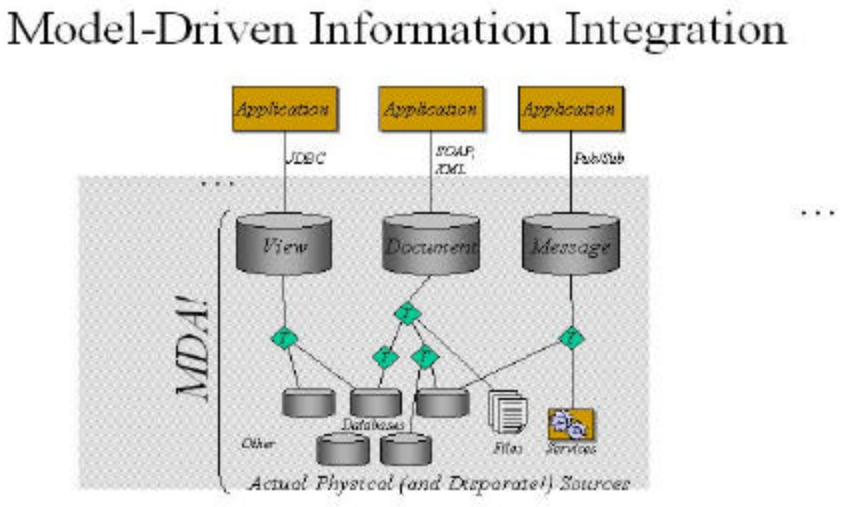


Figure 6: Models Provide Views and Access Paths to Information Sources

One of the most significant services that can be provided with a federated data management model (as shown in Figure 6) is that the business line users, the

consumers of this data, can be provided with an integrated view of information from their point of view. It can provide them with an access method and access mechanism that meets their needs and hides the complexity from them. Actual sources of the data can still be used. There are no copies that are needed. This can provide an opportunity to model the integration with real results. Depending on the frequency those integration steps, they can be made into a canned process or separate stories or even cached files in a highly optimized case. You can use the virtual data base access to model the integration and to define the relationships and transformation functions that are needed. This can support the dynamic and often near-real time demands for new types of data access that can be integrated based on creating new models or more likely pulling pieces of models that have already been designed. New skills in virtual model building will be needed that capture the knowledge of the data and more importantly the analysis skills to understand the information needs of the business community.

Both external service access interfaces and the interfaces within the many layers should be based on open-standards. The interfaces to the "virtual" entities should be standards based open standards and include SOAP to support a web services architecture and JDBC/ODBC to support a client server model. More details on the open standards that apply are defined in Appendix A while the services capabilities are described in Appendix B.

3 Concept of Operations Overview

This section describes the recommended Concepts of Operations to implement the Federated Data Reference Models. This is a high level overview that provides some example technical architectures that can be implemented to accomplish information sharing.

3.1 Challenges and Issues with Information and Data Management and Sharing

Many of the government agencies have a strong focus on information, content, or data issues and related set of improvement initiatives. The immaturity of information and data architectures and the poor linkage with Enterprise Architecture are introducing much confusion and results in issues that are not being addressed in a consistent manner. It is evident to us that there are several challenges that need to be addressed to ensure successful implementation of the Federated data reference model. We have outlined these challenges and issues in Table 1 along with some of the key elements of an approach to address these.

This paper does not provide detail on all the approaches cited in the table, but instead concentrates on two particular items defined in the FEA-PMO DRM initial vision:

- "A framework to enable agencies to build systems that leverage data from outside the immediate domain."

- "A repository that provides multiple levels of granularity to satisfy the re-use of data schema from multiple stakeholder views."

We have taken a strong implementation and action oriented focus around those specific elements of the overall enterprise data management and data architecture approach.

Table 1: Information, Data and Content Issues

Issue	Impact	Approach
Enterprise Information and Data Management is step-child of Enterprise Architecture	Often not considered until late in the development project and when it is addressed it is not addressed from a business perspective but from detailed information, content, and data stand point.	Provide strong ties between business needs and the information, content, and data from a "business-data value" perspective that is linked to the performance reference model elements. Treats "information, data, and content" at a conceptual level and participate in the development of vision. Define a Business-Centric Data Modeling approach.
Data Modeling is known as being dirty-nasty and complex.	It is avoided because what is done is logical or even physical modeling and not conceptual. It is done too late and problems occur.	New skills and mindset is needed. Tool support will help along with a combination of top-down and bottom up approach.
Many-many formats exist.	This creates an inability to "understand" the commonalities and differences between data.	A higher level approach to model driven architecture based on the use of open-standards is needed. Standards must provide a migration path and way to translate and link the many types of data into a higher level abstract approach.
Each container or collection of data is protected by its owner.	Centralized versus distributed control battles rage.	Federated approaches can provide a shared method of ownership.
Data is connected to an individual system.	Data is owned by the system and not really	Data or at least higher levels: containers or

	connected to the "enterprise" or the business owners within the organization.	collections of information, data, or content must be owned by the organization roles, connected to the systems and business context and connected to the locations to receive and use the information, data or content.
Data modeling is at the design level – rather than at the architecture level	Information architecture resides at the Enterprise, provides for a much needed strategic perspective; where as modeling tools to date have had a tactical focus on software development	We need an information architecture which to base our data modeling. We need to view information from an agility model as well as from a connectivity perspective. Our Government requirements need to be addressed from strategic viewpoints.
Metadata and modeling are not well understood and they are difficult to define in business terms.	A severe business to technology gap exists and the result is to stop communicating.	A business centric methodology to provide guidance and data value propositions must be defined and used within a set of business cases for improving the data quality, data and information sharing, and why metadata and models are a critical element.
Emerging technology for information integration has many proprietary approaches.	This reflects the stage of the technology and desire by many companies to strategic try to lock the buyers in.	Ask about standards; foster an open-system standards approach first recognizing there may be gaps. Create an information, data and content standards profile and expected standards. Request a statement on the strategic plans of vendors to support reasonable open-standards efforts.
My business keeps changing – organization, interfaces, processes, etc.	Disconnect between architecture and implementation	Enterprise architecture and models require the use of Choice points for declarative change within

		models and connections to changing trace to implementations eliminating shelfware aspect of architecture products

Process and Activity Model Driven Data Integration

Four of the basic elements of the process and activities needed to create a data and information model are depicted in Figure 7 below. These can be used as the basis for a practical integration engine within business line hubs; as described in the business line architecture and implementation paper (IAC Business Line Architecture and Implementation White Paper – March 2003).

The focus is on reuse and consolidating and integrating the activities that already exist. Once an initial set of valuable resources have been collected by a government wide initiative into a set of business lines (such as those initially selected in financial, HR, benefits, health data monitoring, criminal investigation, and data and statistics, the publication and presentation of the "content" via web services), a set of training and outreach activities can kick start the reuse and commonality effort. There has been a grass root commonality effort for years among both the government and commercial world and now the technologies, early results, and business case can be made to move forward.

The processes and activities as shown in Figure 7 can:

1. Leverage a number of initial data element efforts from projects such as the Justice Global Program, the NEDSS CDC Health Data Monitoring effort, the e-Vital and the eBenefits etc. will create an initial set of government – wide core element files.

2. The second step is to define the use of taxonomies. This has been a recent topic within the XML Working Group (April 17, 2003) and an approach to integrating this topic can be reached by the XML working group.

3. Look for commonality by doing a common comparison across the business lines with a focus on the tasks that are being performed. This is often called semantic integration and can often use ontologies which are formal definitions of terminology in given areas such as those in medicine such as UMLS from NIH National Library of Medicine.

4. Evolve to a federated set of Registries and Repositories that can be linked together as shown earlier in Figure 3. Eventually, the federated registry and repository will have to provide for active data management along business lines. Each of the registry-repositories can be part of a business line hub that can manage such critical functions as the data dynamics to

receive notification of changes to support data recovery and continuity or operations and to support translation between business partners.

5. By focusing on business line data models with common core elements, the agencies can keep the autonomous ownership of their data and "consolidate and integrate" when the business drivers are there. An incremental approach can be used as discussed more in detail in a later section.

A set of parallel activities should also be considered for security and privacy consideration that include looking at the data sensitivity and impact, looking at common security patterns of protection and "anti-patterns" of common threats and the type of security and privacy solutions that are involved in a federated approach. These issues will be addressed in more detailed in a companion security and privacy paper.

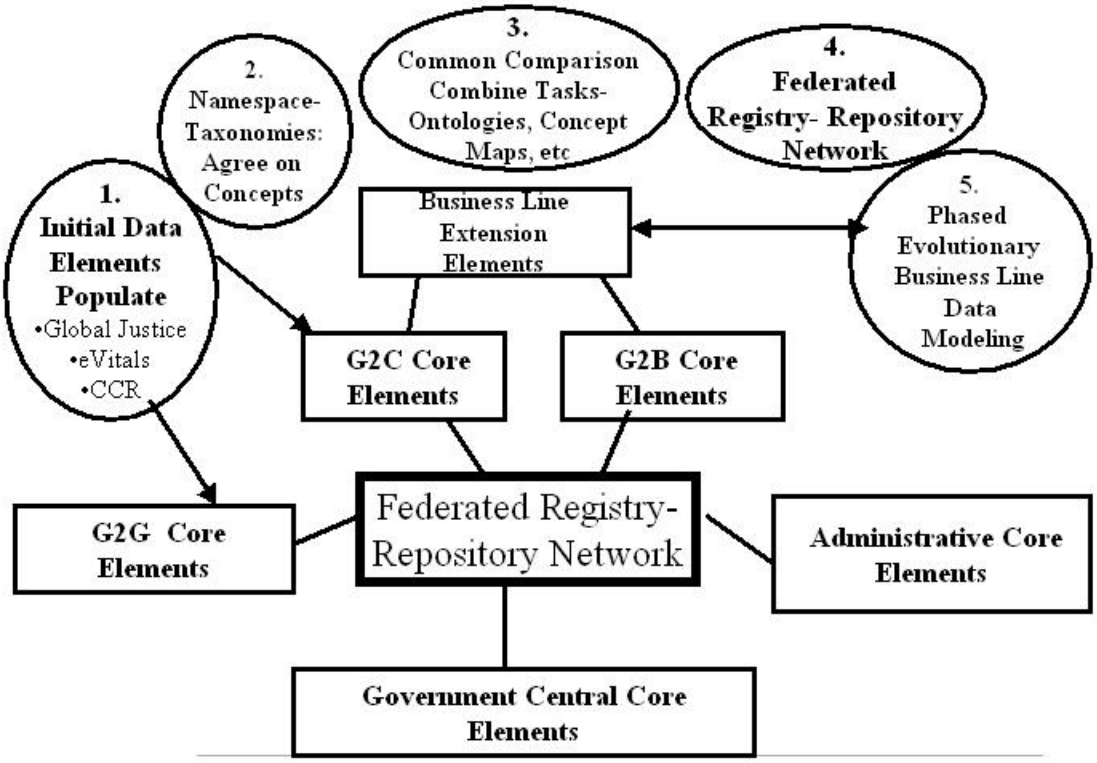


Figure 7: Process and Activities to Create Federated Data Environment

Schema Evolution for Local and Global View

One of the critical elements in very large databases is the ability to evolve the schema and keep track of changes that impact the local or global views. Any truly global view of all government information, data, and content would be overwhelming. The linking of federated data, information, and content models to business lines makes the effort implementable with actions done within each business line and a "small" central function to define the core elements for the government, G2C, G2G, G2E, or other appropriate common elements. A basic set of seed common elements can be defined based on existing projects while leveraging and participating in efforts such as the Universal Business Language, common models such as the Resource-Event Agent, or the Universal Data Representation. A collaborative tool linked to a registry and repository can be a great benefit to this evolving process. The creation of business line data elements (shown in Figure 8) can leverage the reusable core elements for common government needs or types of delivery mechanisms such as G2G interfaces and evolve with the collaboration based on the data representation among business line community partners. Alignment will take some time within each community but collaboration and leveraging the early adopter experience will be critical to the success of this challenging effort.

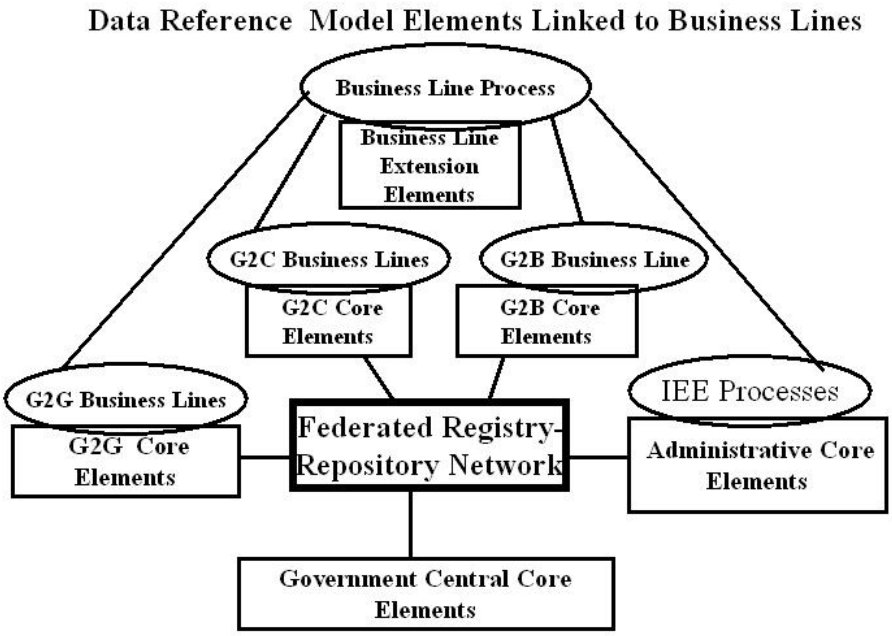


Figure 8: Schema Evolution with Core Government Elements and Extensions

3.2 Technology Concepts, Standards and Readiness

Technology Opportunities: To address some of the issues and challenges presented in Section 2, there are technologies that are ready today and there are technologies that are emerging and those where the concepts have not been fully formed. The needs for improved information and data management with the Department of Defense, Homeland Security and across the business lines are the most challenging. Many of the challenges will be addressable, but many of the cross-domain dynamic data architecture elements are still on the research drawing board and have not addressed the scale and performance that the government needs. We have seen many Enterprise Architecture initiatives that have not risen to the challenges of information, content, and data management. Most of these have failed due to lack of connection between their Enterprise Architecture and their business goals and drivers. With emerging technologies, a systematic business driven approach is needed. We have used a Technology Management process that is Value Driven and builds on the experiences of others and linked to the needs and expertise of the "thought leaders" within the market space.

Flexible Building Blocks: The government- Federal, State, and Local is a large market space and presenting their challenges to the information, data, and content management organizations- university research, standards bodies, and companies can offer great benefits for closing those gaps. We have first defined the current landscape and directions. Then, we defined a set of gaps that are critical to meeting the challenges and finally the key concept leaders that should be "tracked" and collaborated with. No one can forecast the future but we can know the

relationships and "effectual reasoning" that derives the next wave of capabilities. Many of the ingredients appear to be ready but the process and integration elements especially the "market" push from a weak commercial market may not support the level of innovation that is needed without the government being an early buyer and innovative procurer.

Key Emerging technology elements include the following:

- Conceptual Information Modeling and Meta Data Management with Open Standards
- Universal and Domain Specific Syntactical and Semantic Languages
- Service Oriented Architectures, Web Services, and Standard Message based Delivery technologies
- Evolution and dynamic architectures
- Use of rules, constraints, and embedded reasoning
- Model driven and generative programming to allow working with high levels of abstractions.

Conceptual modeling that is tied with the business needs and users understanding of "information" and the creation of domain specific or visual notations that relate to the end users are having a renewed interest. The possibility of having Conceptual Level Query languages built on top of Models and metadata is also prevalent in current thinking. A set of open standards and research initiatives reflect this high level abstract thinking that is closer to the way the end user wants to engage information as opposed to the bits and bytes of technologies. User oriented models have to bridge the gap to the data and information.

Model Driven Architecture Standards: The most universal language may be XML and its ability to create families of languages along with query, transformation, path, pointing, and exchange mechanisms. XML is permeating every part of information, content, and data management but must build bridges to the existing data, information, and content resources through incremental step-by-step transformation to become "mainstream" entities.

XML along with SOAP and the XML-based security standards are standard defined by W3C and OASIS that are critical to data management and interchange. They can be used to define industry and government specific standards for interchange of data between diverse systems. It is important to understand that XML is not a modeling framework and is not an integration framework; it is a messaging format that is useful for interchange and to reach data interchange agreements between partners.

The modeling architecture we propose is based on standards from the Object Management Group (OMG), which promulgated CORBA and UML ten years ago and is now concentrating on metadata management and modeling standards. Currently OMG is articulating an overarching theme named Model Driven Architecture, which leverages past success in modeling and middleware standards and incorporates some newer specifications regarding metadata management. Four years ago this consortium created the Meta Object Facility, a specification for managing metadata for any sort of system within a single framework, the core of this architecture is a modeling framework derived from UML and now based on MOF.

A second important standard defined by OMG is XMI, XML Metadata Interchange, which specifies the manner in which metadata can be exchanged between systems and tools. A critical feature of XMI is that is carries a metamodel description along with the metadata.

Example Scenario Using MOF models and XML based Web Services

The Department of Homeland Security needs to be able to find arrest records for individuals. An example of such as federated data management environment is shown in Figure 9. The problem is that each agency, state, and county keeps arrest record information in a different sort of system: a document management system, relational database, or a flat file. The solution to this problem is to first agree on an "exchange" format and the best candidate for such a format is an XML schema. The Line of Business owner will need to develop a schema that indicates the attributes and structure for the document "Arrest Record." Each individual jurisdiction may or more likely may not have all of the attributes indicated in the schema available in the various systems that they use in their criminal justice operations.

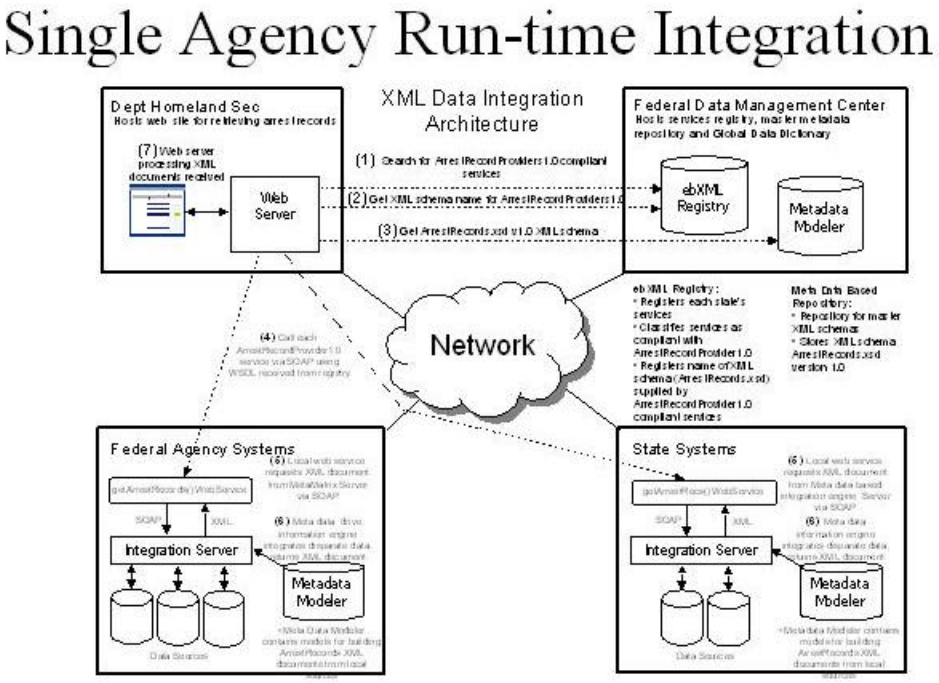


Figure 9: XML Data Integration Architecture

Once the schema is available to be viewed and retrieved by each jurisdiction, they map their physical systems to the schema using virtual metadata – the virtual entity created in this case is the same as an XML document that complies with the schema. After the mapping is complete, the document is reported as a Web Service (component) and can be retrieved by any properly authorized entity. These documents are then reusable components that are available to be used in applications other than those for which they were originally designed and implemented.

4 Recommendations

Given the need for information sharing and business line convergence, we recommend that a high priority be given for implementing Federated Data Model immediately. Information sharing requires that the "current" state of information systems be available: what, where, when, and who owns it and who accesses it must be a top priority. For example, this is critical to the new Department of Homeland Security and high priority government transformation efforts. Currently, Standards and mature technologies are available to leverage the information sharing opportunities. Numerous government agencies have been

investigating and implementing these approaches and we recommend that the federal government take a leadership role in making the information sharing a success.

4.1 Adopt a Model Driven Architecture for Federated Data Management

4.1.1 Modeling for Understanding and Plan for Improvements

The complexity of the Federal government information and data management problems requires a new approach and demands that the best tools be made available to the government–wide community of information and data analysts including state and local organizations.

- Adopt standards and processes whose goal is to build models of an ever-growing number of information management systems across the federal government
    - The modeling standard should be MOF from the OMG
- Construct a Federated Data Dictionary
- Use a MOF metamodel to contain the dictionary so that datatypes may be related to attributes in models
- Create centers of competence that can generate the virtual views from collections of physical models required to support the Line of Business

Based on the understanding of what exists, a set of forward engineering modeled on business-driven needs can create a set of business line focused Information and Data Management improvement plans.

4.1.2 Fit the Situation with Levels of Model and Management

The government has organization elements with limited information sharing needs at one end of the spectrum and some of the most complex and tightly integrated needs on the other. One size will not fit all the data modeling and management needs. Data modeling and management depend on the situation but there are some basic “asset management” steps that are needed by all with federation or joint models being high level and need more powerful tools and approaches. Conceptual we have tried to show this multi-level range of models in Figure 10.

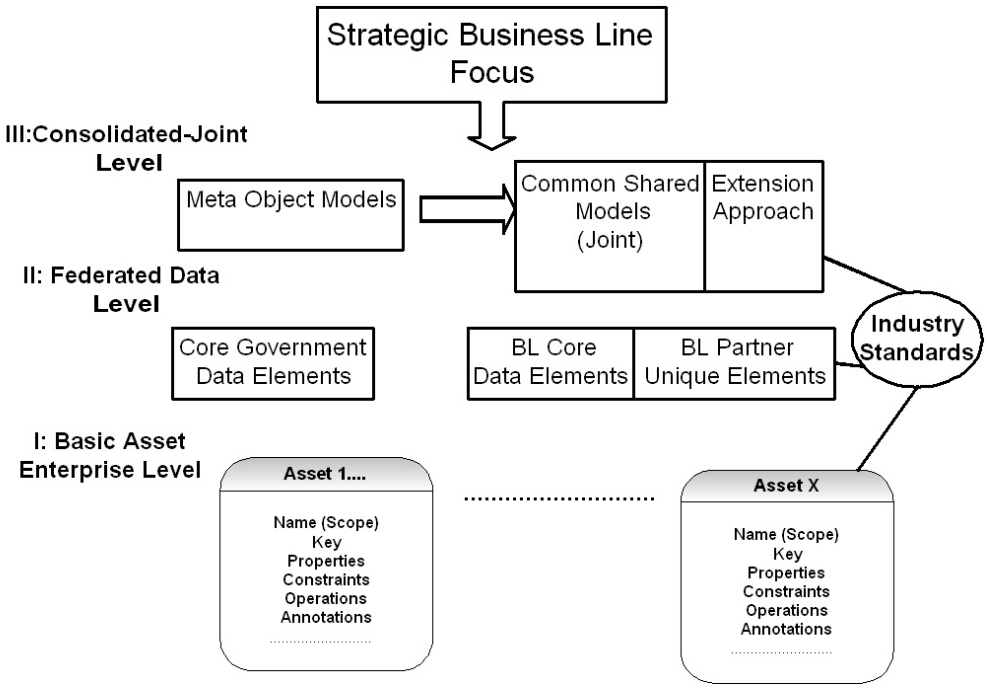


Figure 10: Levels of Data Modeling and Management

The basic asset level can be used across government while the federated data level focuses around business lines and represents a level II set of model-based constructs. The most complex and tightly integrated represent the few consolidated and jointly developed set of models at level III. All the efforts must be driven by the situation and needs.

This situational approach is also reflected in the use of standards or definition of common practices to be used government or business line wide. The standards again depend on the situation. A set of standards can be integrated into a profile that can be mapped to type of needs or complexity of the situation. Figure 11 shows a set of profiles that map to the three level structure in

figure10.

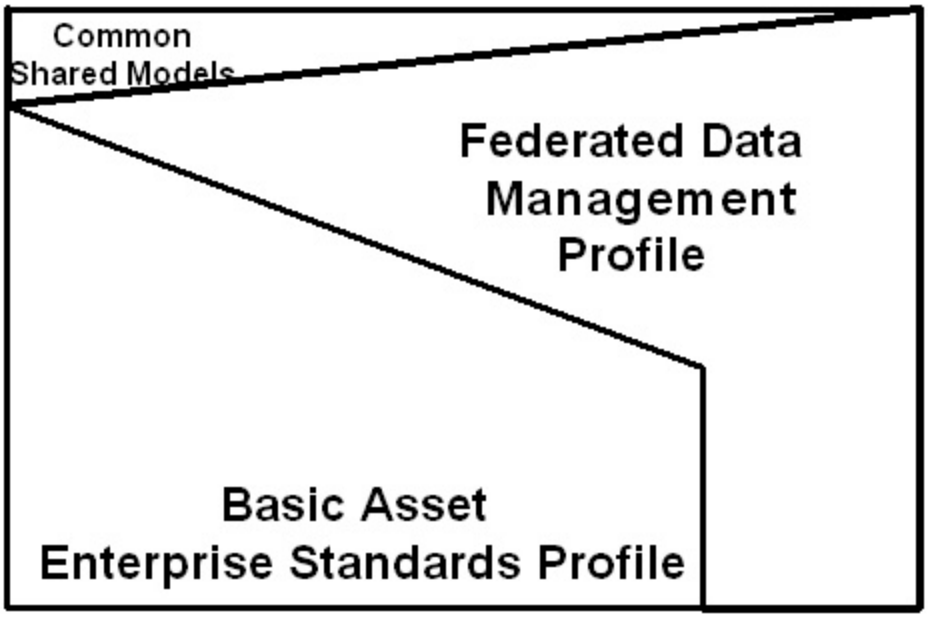


Figure 11: Data Management Standards Profiles Fitting Your Needs

This paper is focused on the level II federated data management levels and related standards.

4.1.3 Model Driven Information Integration

The models developed can allow the information and data to be integrated along business lines. It can have immediate benefits in areas like Homeland Security.

- Implement an integration engine that can use the models developed as part of the "understanding" activity to integrate information from disparate sources

- Manage the integration engine as part of the competency center

An Emerging technology subgroup working closely with Web Services can define a set of integrated web services to connect, discover, access, and provide different modes of information and data services. This group should work to assure that the integration activities are linked with other activities such as registries and information exchanges discussed in the Interface, Information, and Integration IAC paper and that a pilot action plan on a Homeland Security related project and

an e-government project such as the Small Business One Stop be considered to validate the concepts from the information sharing set of papers.

The Federated Data Management approach should be tightly linked with registry standards such as ebXML and UDDI and integrated with the ongoing registry-repository projects lead by NIST. OMG and OASIS should be encouraged to "collaborate" on the standards effort and the government and industry representatives supporting government should be active participants in the OMG and OASIS Government task force efforts that the IAC has encouraged.

4.2 Develop Information and Data Policy-Driven Services

The Federal Government has begun information and data quality improvement initiative within OMB and a number of agencies have ongoing data quality efforts underway.

With the proposed service oriented architecture a consistent delivery of information and data policies can be done by applying three simple steps:

- Define a unified set of information and data policies
- Provide those information and data policies in assertion and constraint based XML language
- Provide a set of policy enforcement management

Information and data policies can be delivered in a more consistent manner if they are managed by the business lines, but with a set of well-defined services.

We need a unified data policy that encompasses rules and regulations in the following areas:

- Data security
- Data privacy
- Data Corruption prevention
- Customer data protection
- Access rights

4.3 Establish Information and Data Migration Services: Extraction, Translation, and Loading

A unified program should be established at LOB level to manage the data migration activities across LOB/Sub-Functions including the following:

- Define strategies for synchronous and asynchronous data movement across LOB/Sub-Function information and data centers

- Utilizing commercial data movement and transformation tools efficiently

The Federal Government should facilitate a government-wide data integration program to support the information and data integration across the business line. This program should include the following areas:

- Defining standards for developing and managing data models at LOB/Sub-Function level

- Defining strategies and approached for supporting Shared Information and Data across LOB/Sub-Functions

- Defining strategies for implementation of the Information Value Chain process at LOB/Sub-Function level to ensure Government, Citizens, and Business Partners efficiently can collect information, store, query, distribute, analyze, act, and learn from it

4.4 Enable Information and Data Stewardship: Registration and Notification Support to the Owners and Users

The Federal Government should establish a unified process for managing Data Ownership at LOB a Sub-function level including the following:

- Data Stewardship at LOB/Sub-Function level must be identified in order to understand who is the creator or originator of data and who uses it

- Data Stewardship program should be established at LOB/Sub-Function level to define roles and responsibilities of owners, custodians, and stewards. The primary benefit of this program is the elimination of the unnecessary duplicative data sources across LOB/Sub-Functions. Also it may assist lines of business to identify overlaps and gaps

4.5 Advance Data Dictionary and Metadata Management

An important benefit of the Model Driven approach is the generation of a Federated Data Dictionary using a MOF based Datatype metamodel.

- All data entities related to the government information need to be defined in a centralized repository that is accessible to all government's organization, citizens, and government business partners

- All business rules must be stored into a centralized repository. This information must be shared across government agencies and its customers

- LOB/Sub-Functions must be responsible for maintenance of the Information and Data Repositories in order to reduce the maintenance cost and promote information sharing across agencies

4.6 Invest in Information and Data Administration: Tools, Skills, and Training

Data Centers at LOB/Sub-Function level should be responsible for data administration activities including the following:

- Model Development and Management
- Data and Metadata Change Management
- Data and Metadata Version Control
- Implementation of Data rationalization
- Implementation of Data standards

4.7 Lead Information and Data Standards: Selection, Certification, Extending, and Requesting Exceptions

Data Standards should include the following activities:

- Defining Data Naming Standards
- Definition and Maintenance of Lexicons and Acronyms at Government level
- Define rules for Data rationalization at government level
- Defining processes and procedure for data administration at LOB/Sub-function level
- Defining Interface Standards

5 Summary and Conclusion

Context: Federal Government can be more efficient and more cost effective if it changes the paradigm from Agency-Centric to LOB and Citizen Centric. In a Citizen Centric environment, agencies, citizens, and their business partners can share information and data that is collected once and distributed anywhere uniformly. It will be easier to identify and eliminate the duplicative areas as well as provide capabilities to identify gaps and opportunities for improvement.

Summary: This paper outlined the concept of the Federated data model, developed some innovative models for an extensible framework and provided concept of operations to implement such a reference model. This data reference model is tightly linked to FEA PMO reference models. This paper also provided some concrete examples of how information sharing and business line convergence can be accomplished. Then a number of recommendations are made to advance the model for Citizen Centric government.

Recommendations: The transformation to information sharing will take a number of concurrent and integrated steps as shown in Figure 12. This will take investment and a few that crosses many government agencies, it cannot be done without investment and cross-government commitment but it is critical to the success of these Transformation efforts.

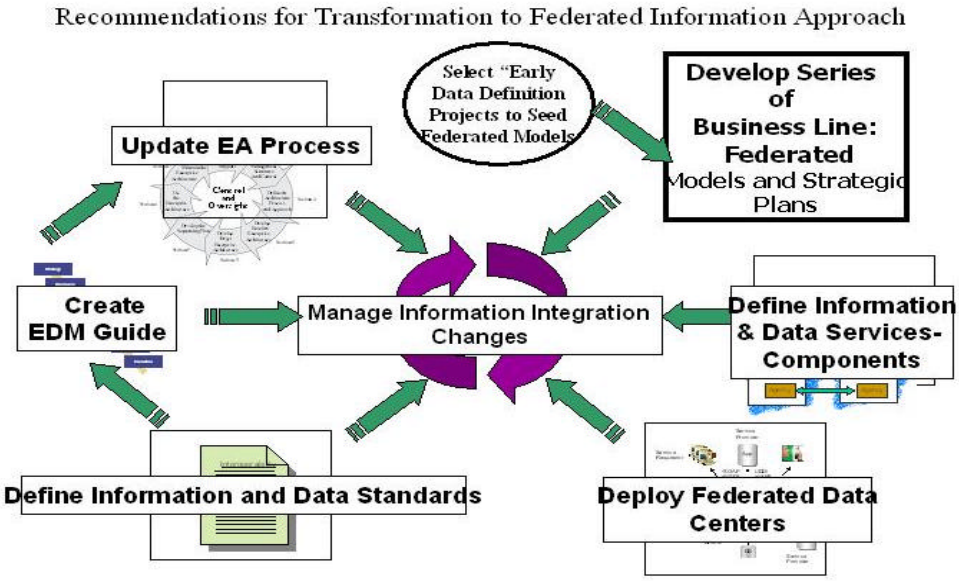


Figure 12: Series of Actions needed to enable Business Line and Data Integration

Federated Data Management provides standards and tools for developing Information Models and Data Models for the Federal Government at LOB/Sub-Function level in a unified manner and based on the cross-agency standards. Federated Data Management Centers will create collaborative environments for Federal Government, citizens, and business partners to share information and data. The utilization of Information Value Change process is an integrated part of these Centers.

Federated Data Dictionaries can be used to inventory information entities across agencies at LOB level. The process used to generate Federated Data Dictionaries should be based on OMG's MOF modeling and metadata management standard. These models can then be used to drive a component based integration framework.

Establishing Federated Data Management Centers at LOB/Sub-Function level will reduce the costs of operations and maintenance. All centers will operate upon uniform processes and will follow the standards.

Appendix A: Enabling Technology and Key Trends

Model-Driven Architecture: Status and Deployment

Model Driven Architecture (MDA) was developed at the OMG based on a decade of experience in creating portability and interoperability standards that have been successfully deployed in enterprise and mission critical applications. It is based on a fundamental pattern comprising of Platform Independent Models (specification, code etc.) that are mapped to Platform Specific Models using standardized mappings that enable automation of the mapping process. It is observed from experience that this pattern is repeatedly applied recursively in typical development and deployment environments, thus opening up the possibility of creating tool chains covering the entire process that can be populated by tools, from many different tools vendors, that can participate in the chain through standardized information exchange formats thus significantly automating the process of system generation and deployment and reducing cost of deployment.

Early modeling language and model exchange standards like UML 1, MOF 1 and XMI 1 have now matured and they have been used successfully in large systems. An evolution from this early effort was the Common Warehouse Metamodel (CWM) standard that has been used successfully by major Data Warehouse vendors in their products. Thus, there is considerable real world deployment experience to justify consideration of these standards as the basis for a standardized architecture for the Federated Data Management problem. An evolutionary step towards the completion of the UML 2, MOF 2 and XMI 2 standards is about to be completed. This builds on the experience to fill in known gaps in the earlier standards and making it cleaner to do some of the functions needed in data management.

Model Driven Architecture Simplifies Integration by Separating Architecture from Implementation

The Object Management Group's (OMG's) Model Driven Architecture (MDA) helps protect organizations' data and software investments by capturing higher-level semantic information; whether it be about data sources, software modules, business logic, or business processes and their appropriate interactions -- in re-usable models. The MDA can significantly simplify the integration of disparate computing artifacts of all kinds, as the architecture separates the models that represent the artifacts from the particular forms (or instances) that those models take in a particular system. The MDA represents an advanced approach to software design based on models created using languages such as the Unified Modeling Language (UML). As business requirements change or evolve, components of model-driven applications can easily be replaced, as long as they

are standards based. Adherence to open standards ensures that there can be no vendor lock-in. Models use a set of metadata -- the data describing the structure and characteristics of data or program elements. For example, customer data might include a first name, last name, address, and phone and customer numbers. Metadata describing customer data would specify whether these elements are alphanumeric or numeric, the maximum length and format of the phone and customer numbers, and so forth. Although the way these are represented may vary from data silo to data silo, a common model can allow for automatic integration across the silos. Both commercial products and open source implementation exist. The open source Metadata Repository (MDR) exists within the NetBeans[tm] open source project (www.netbeans.org). The MDR supports the Meta Object Facility (MOF), the technological foundation of the Model Driven Architecture. This makes the NetBeans platform the first of its kind to support MDA. The MDR uses the Java[tm] Metadata Interface (JMI), the standard for expressing metadata in the Java language. The MDR also supports XML Metadata Interchange (XMI), the OMG standard for metadata interchange. The MDR makes it easier for developers to support other programming language without extensive extra programming, and to write NetBeans-based tools that interoperate with standards-based modeling tools. The Sun[tm] ONE Studio developer products (formerly Forte[tm] tools) are based on the NetBeans platform. The NetBeans community is pioneering model-driven development by providing tools for other developers, and as the architectural foundation for the platform itself.

Both leading companies and start-ups lead by thought leaders support model Driven Architecture. An example is the effort that HP has made in the definition of modeling standards activities since the early days of UML 1. HP led the UML 1 standardization process to successful conclusion at the OMG. HP also played a leading role in articulating the vision of MDA, and continues to play a significant role in evangelizing its use and evolution. Today, companies are focused on their customers to more efficiently deploy and manage their IT resources and the increase in adaptively. Company mergers and acquisitions in the commercial world and the need for information sharing and reorganizations such as the Department of Homeland Security will require the successful use and interoperation of tools and software both within the large software supplies and among them and the network of partners, for which standards based interoperability is a critical success factor. As part of the evolution of organizations more Adaptive Management Infrastructure at HP and similar initiative in other organizations the large vendors are exploring ways of exploiting MDA related standards in the associated tools through our participation in the Eclipse Consortium along with the participation in standards organizations such as OMG and OASIS. The government can be a positive influence on this market by working on these open standards and influencing the direction of the market to more adaptive and interoperable tools for data and information sharing. We expect MDA architectural principles and standards to play a significant role in enabling vendors to provide more cost effectively deployed Adaptive Management Infrastructure in a heterogeneous environment comprising of operating system and devices platforms from multiple vendors.

Information and Data Delivery Web Services

The data modeling activities must be a key precursor step to the delivery of data. We have begun with the end in mind and looked at the delivery functions and worked back to the technology needed to model and understand the data, via models. We look at the mechanism of delivery via web technologies. Mechanisms of delivery to the citizens, to businesses, between government organizations must be clearly defined. In this appendix we examine the technologies needed for model-driven architecture and model-driven integration and the elements that provide for information interoperability.

Information interoperability is really a portfolio of technologies that must work together and to improve the internal efficiency. Information, data and content modeling must be integrated with the delivery mechanisms. The "delivery mechanisms" defined in the new Business Reference Model 2.0 must be connected to the Information and Data Reference Model for both understanding what information, data and content currently exist, for planning the "to-be" information, data, and content needed and the mechanisms that will be used to meet the business needs. There can be a number of "delivery mechanisms" that are used but our focus has been on the applicability of web services as a primary mechanism.

XML is the format in which information will be described when accessing the federated layer. There are two approaches to the structure of the XML documents that can be retrieved. Either a schema can describe them or they can be defined ad hoc. It seems that most of the federal apparatus has proceeded down the path of pre-defining a collection of schemas to fully describe the data requirements for a line of business. For instance the Department of Justice has recently completed JXDD 3.0, a collection of XML schemas that defines the information requirements for criminal justice. At some point in the future it should be possible to retrieve any information concerning criminal justice by asking for XML documents whose content and structure is defined by these schemas. The other option would be to determine what data is available from physical systems and use Xquery through a translation layer to retrieve ad hoc documents.

In the first case, a significant challenge arises in the process of defining a set of XML schemas to describe the data for a line of business, criminal justice for instance. It is exceedingly helpful to have a description of the physical data stores that house the desired information, including the data elements available in each source, the schema of the source and some other characteristics of the data management system, during the design of the XML schemas. With complete knowledge of the characteristics of the physical information stores, XML schemas can be designed so that the mapping between the XML document and the schema of the physical can be as simple as possible, and the data set described completely in the XML schemas.

The real world however, is that no matter how much care is taken to make the XML schema conform to the schema of the underlying physical sources, they will almost never map completely. The largest impediment to implementing a web services infrastructure is this issue of mapping XML schemas to non-XML data sources in a way that facilitates real-time access of correctly structured XML documents. This mapping information can be specified directly in code or it can be specified in models and the models can be used to drive the real-time production of XML documents.

The key concept is that a complete set of metadata describing the underlying sources and the metadata describing the XML schema can be manipulated in the same modeling space and the mapping and transformation information can be specified as components of the models.

These "mapped" documents can then be described by WSDL, and registered in a UDDI registry so that they are a web service. This essentially provides a web service for data access.

A.1 Information Interoperability Service Overview

In order for business-oriented services to provide value within an organization, they require information and data upon which to act. The provision of this data, as well as the related metadata, as part of the overall architecture of the Info Fabric, is called the delivery layer. This layer provides the standard means of communication among different software applications, running on a variety of platforms and/or frameworks, for the purpose of de-coupling the services layer from the physical realities of the data it must process. The services layer requires a common set of description, connection and access capabilities for information and data that can be shared by the services that will use it.

This common set of description, connection and access capabilities has previously been attempted through a variety of approaches that have met with varying degrees of success. Some of these approaches have been implemented within the federal government on a large scale and in a production environment, while other approaches have only been implemented in pilot initiatives. These approaches, along with their relevant standards and specifications, are described below.

A.2 Data Dictionaries

A data dictionary is a subset of metadata that is explicitly exposed (published or packaged) for shared use for the purposes of discovery, reuse, analysis, or general information. A data dictionary is broader than just datatypes - it also includes data elements, and compositions of elements, including instances of both simple and complex datatypes. In many instances, data dictionaries have been implemented from a data element viewpoint (as instances of simple datatypes).

The ISO 11179 standard has been utilized in the implementation of data dictionaries. It was finalized in June of 1998 and provides guidelines for the description of individual data elements, and is defined in six parts as follows:

ISO 11179-1 Framework for the specification and standardization of data elements: Provides an overall view of how data elements should be annotated and classified.

ISO 11179-2 Classification for data elements: Defines an ontological classification mechanism for the unambiguous classification of individual data elements.

ISO 11179-3 Basic attributes of data elements: Provides a list of metadata that should be collected for data elements. These include information about the value itself, as well as information about the meaning of the element.

ISO 11179-4 Rules and guidelines for the formulation of data definitions: Provides best practice definition approaches for individual data elements.

ISO 11179-5 Naming and identification principles for data elements: Provides some standardized mechanisms for the unique identification of data elements.

ISO 11179-6 Registration of data elements: Describes a methodology for the registration of individual data elements.

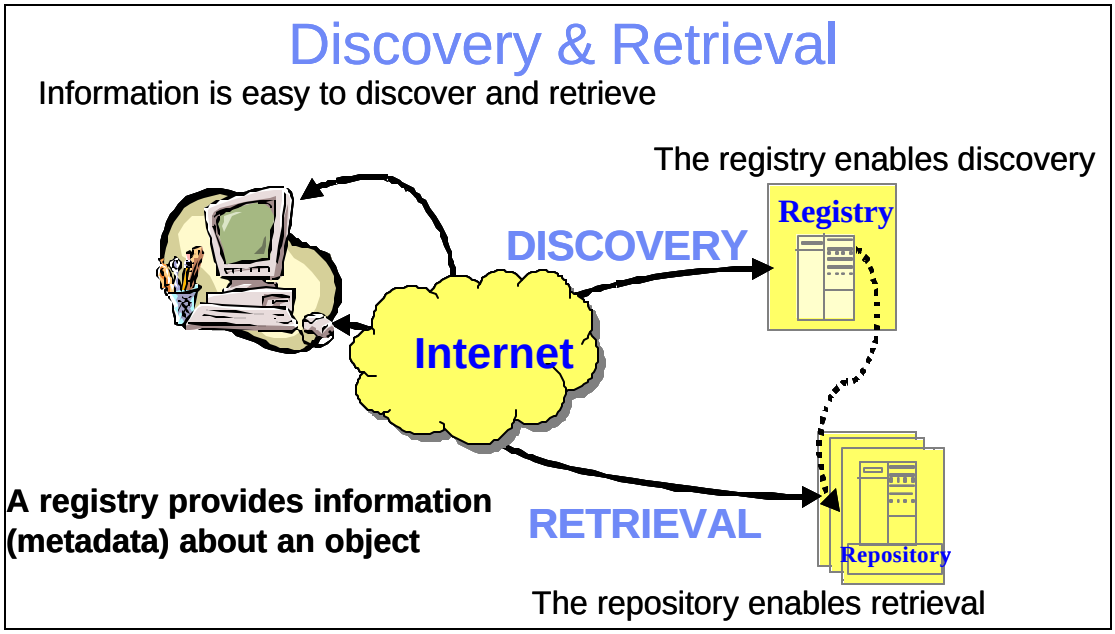
While this standard provides an approach to the classification of data elements, as well as a mechanism to define the allowable value domain for each element, by itself it does not provide enough information to create a robust data dictionary that can interoperate with registries and other components required for business-oriented services. Some of the features lacking are a mechanism for the association of data elements into structures, and a rigorous way to retrieve associated elements programmatically.

In order to fully interoperate within a services-oriented architecture, a data dictionary needs to encompass more than just datatypes - it includes elements (simple and complex), and relationships. A data dictionary is the result of "packaging" or "publishing" metadata with the explicit intent to share it for a purpose, with different data dictionaries to be published for different end-user communities. A data dictionary would need to be maintained and managed in electronic form, but is just as likely to be "consumed" by end-users as a reference document in hard-copy form. End users of data dictionaries include business analysts, data analyst, architects, developers, and managers, and the uses of data dictionaries include discovery, reuse, analysis, application development, and general information.

A.3 Registries and Repositories

Registries provide a common location where metadata about specific data elements, structures, and/or service interfaces can be registered by organizations

that want to share that data. With a registry, the emphasis is on sharing between organizations, and the focus is the enabling of e-business/e-commerce. There are several standards related to registries, including ebXML, UDDI, and the JaxR API. The purpose of a registry is to enable dynamic and loosely coupled collaboration, while at the same time formalizing how information is to be registered and shared. Registries are typically used beyond the scope of a single organization or agency. The Discovery and Retrieval functions are shown in Figure A-1.



Ref: OASIS ebXML Registry Standard, Katherine Breininger, 20 Jan 2003
Figure A-1: Discovery and Retrieval Using Registry and Repositories

A registry may contain a number of objects, such as XML schemas, sample interface documents, descriptive information on schemas, descriptive information on structures that comprise schemas, and web service descriptions (WSDL's). They may also contain metadata about its contents, such as the submitter, lifecycle status/dates, content classification systems, a glossary of terms, information on companies responsible for creating content, and contacts at those organizations

Registries and repositories typically work in tandem. A registry typically provides an interface that is used to maintain a repository, and a registry holds a catalog that describes the submitted content in the repository. Meanwhile, a repository typically provides the stable store that holds submitted content. The following list represents the services that a registry provides, relative to the registry and/or repository contents (JaxR is an example of a registry services interface).

Register new objects

Provide metadata for objects

Browse or query registry context

Filter out irrelevant references

Retrieve context of selected items

Figure A- 2 describes the relationship between the registry API, the registry, and the repository.

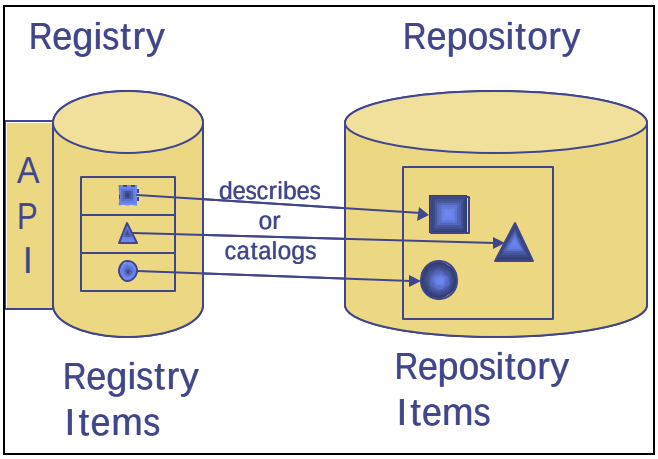


Figure A - 2: Linking between Registries and Repositories

The key capabilities of a registry-repository combination includes the following:

Registry stores information about items that actually reside in a repository

Registry serves as the index and application gateway for a repository to the outside world

Registry contains the API that governs how parties interact with the repository

Items in the repository are created, updated, or deleted through requests made to the registry

Roles are defined within a "yellow pages" for a registry, which includes a directory of the organizations with registry access (and their access privileges), as well as a limited number of contacts for each organization (and their access privileges). The following roles are typically required in the operation of a registry:

Submitting Organization – owns content and publishes to a registry

Content Submitter – User in a Submitting Organization, authorized to submit content on their behalf

Registry Operator – Has special access control and authorization privileges

Registry Guest – Can browse/search data in registry

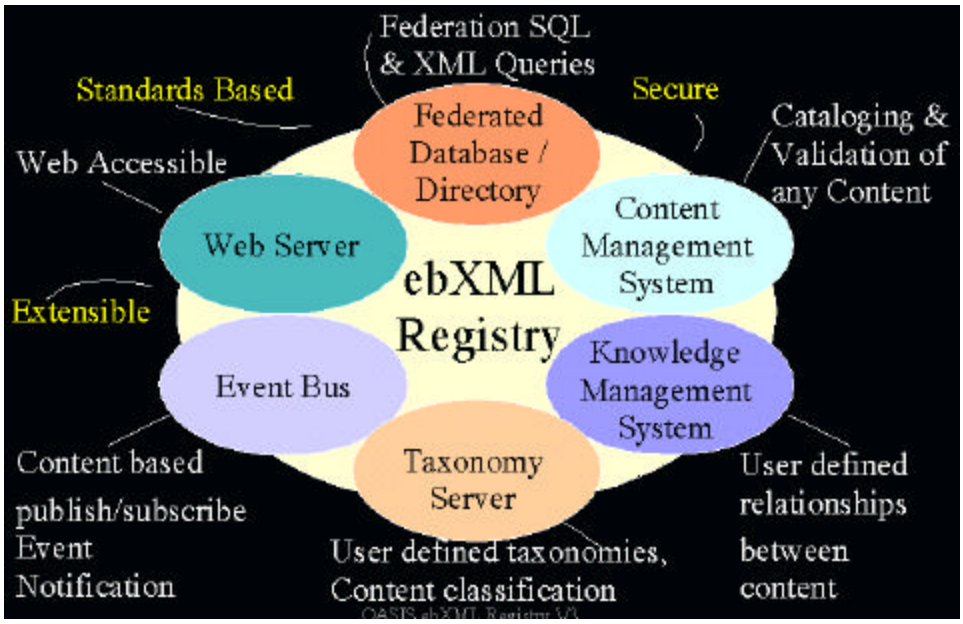
A.4 ebXML

The ebXML standard provides a specification for an infrastructure that supports the publishing, discovery, and management of evolutionary change related to information and data models in a service-oriented architecture. The ebXML mission, as defined by ebXML.org, is "to provide an open XML-based infrastructure enabling the global use of electronic business information in an interoperable, secure and consistent manner by all parties." It is focused on the reuse of business processes and methodologies across businesses and industries, and therefore is much more than just a registry. The features and functions which ebXML supports are described in more detail below and in Figure 3

EbXML 2.0 Features/Functions:

- Includes a registry plus a repository
- Registration and classification of any type of object
- Taxonomy hosting, browsing and validation
- Association between any two objects
- Registry packages to group any objects
- Links to external content
- Life cycle management of objects
- Flexible query options
- Built-in security
- Event-archiving – complete audit trail
- Service registration and discovery
- EbXML 3.0 Features/Functions:
- Cooperating registries
- A registry may cooperate with multiple federations (for the purpose of federated queries, but not lifecycle mgmt)
- Inter-registry relocation, replication, references

- Federation metadata is stored in one registry
- Event notification
- Content management services
- For pluggable content validation, discovery, cataloging
- HTTP interface to registry
- Registry metadata & content addressable via URL
- Enhances interoperability with other types of registries
- Iterative query support (“more”)



Ref: Technical Overview of OASIS/ebXML Registry v3.0, Farrukh Najmi
Figure A- 3: ebXML Registry Overview

While the ebXML standard provides an approach to the registration of many types of objects used for electronic commerce, as well as a mechanism to define XML schemas, it does not support all of the capabilities needed to provide a complete infrastructure for local and global views required for data integration. The ebXML registry standard is focused on the sharing and reuse of XML Schema documents. While it provides extremely robust business registry functions, the scope of registered ebXML objects is not granular enough to support the true reuse of data items and structures.

A.5 Classification Schemes

The classification scheme identified within the context of ISO 11179 and ebXML provides for a number of uses:

Find a single element from among many

Analyze data elements

Convey semantic content that may be incompletely specified by other attributes such as names and definitions

Derive names from a controlled vocabulary

Disambiguate between data elements of varying classification power:

However, other classification schemes may also be required to provide additional capabilities within the context of a de-coupled services layer. These additional capabilities are listed below, in the order of increasing classification "power" and/or complexity.

Keywords

Thesauri

Arranges descriptive terms in a structure of broader, narrower, and related classification categories

Taxonomies

Provides a classification structure that adds the power of inheritance of meaning from generalized taxa to specialized taxa.

Ontologies

Rich, rigorously defined structures and constraints that can provide semantic information to software components

A.6 UDDI

The Universal Description, Discovery, and Integration specification (UDDI) provides a dynamic registry for Web services. Businesses that wish to register their service offerings with a UDDI registry may specify metadata about their company, as well as business and technical metadata about their service offerings. The UDDI registry model is specifically geared to the description of Web services. More granular information, such as the metadata for individual structures within the Web service messages or the metadata for data points within those structures, cannot be provided through the UDDI registry.

A.6.1 Comprehensive Delivery Layer Requirements

A comprehensive approach to a delivery layer, including an infrastructure to support publishing, discovery and evolutionary change, needs to include the relevant components from the approaches and standards defined above. ISO 11179 provides a mechanism for the registration of individual data elements; ebXML's registry model describes a registration strategy for XML Schema; and the UDDI specification details the registration of Web services. In an ideal implementation, all of the metadata for data elements, XML schema, and web services should draw from the same registry or set of registries. A robust registry would allow specific data elements, structures, and datatypes to easily be reused. The registry specifications that exist to date do not directly provide support for the reuse of individual data structures and data points outside of the context of an XML schema. The registry should allow these information components to be individually tracked and versioned, and any other part of the registry should be able to leverage these components in other structures.

In a true enterprise-class solution, there may be thousands of different participants in the information design and registry process. In many cases, there will be a certain level of overlap in the information needs of the various participating entities; on the other hand, many information items will be proprietary to a particular stakeholder. The requirements of the repository that supports all of the relevant registry capabilities described above are significant enough such that no single repository will suffice. The integration of metadata from a wide variety of sources needs to be accomplished in a repository that addresses the following:

A logically centralized view of enterprise metadata, either through a virtual repository or federated repository

Abstraction layer to hide the technical and physical details of metadata storage and management

An industry-standard, integrated framework for a wide variety of metamodels

Access to internal or external metadata sources

A federated architecture to dynamically adapt to the defined scope for any given agency or organizational

Fully featured, comprehensive search, query, and report functionality across the federated metadata repositories

An infrastructure that is able to integrate any kind of metadata source, including both import and export capabilities

Repository access for a variety of end-users/roles via a customizable portal

Access to the repository using open interfaces.

These ideas reflect many of the key concepts described in more detail in reference 10.

(Ref: XML Collaborator: XML Design Collaboration And Registry Software White Paper September 2002, Kevin Williams).

A.7 The Meta Object Facility - MOF

Any attempt to integrate data and information across disparate systems will ultimately depend on discovering the metadata in those systems. There is no method that does not begin with the discovery and analysis of metadata. Collecting metadata is only half of the baseline problem – relating the metadata in one system to the analogous metadata in any other different system is the second and more important problem.

A strong effort, led by a few early government adopters (such as the NSA Middleware effort, and the DoD Single Integrated Air Picture program), has applied the Modeling and Meta Data Management and Standards effort lead by OMG. We have known for a long time that the key to a truly federated data management capability is the ability to manage the metadata from diverse systems. Until recently there has not been a standard based way in which this could be accomplished. MOF and UML 2.0 efforts in combination with XML can offer that opportunity. These capabilities are key emerging technologies that can play a significant role in the interoperability and information sharing elements of the Federal Enterprise Architecture.

Meta Object Facility (MOF) enables all metamodels and models to be defined in a single "language", and since it is a single language, there are no walls preventing the capture of cross-model relationships. MOF is an extremely powerful modeling language that can define many meta-models (relational, object, XML Schema, XML documents, UML, business processes, workflow, etc). OMG has defined many "standard" metamodels, UML, CWM, etc., but the real power of MOF is in enabling the definition of any special purpose metamodel required to facilitate a Model Driven Architecture.

MOF is a layered metadata architecture consisting of a single meta-meta-model (M3), meta-models (M2) and models (M1) of information. Each Meta level is an abstraction of the Meta level below it by describing the format and semantics of the layer below. (These levels of abstraction are relative, as one person's data is another person's metadata.) The purpose of these degrees of abstraction is to achieve a single architecture for constructing models of highly diverse systems.

These models and the relationships between components of the models can then be used to facilitate understanding of a collection of complex systems in a single

modeling tool. More importantly though, these same models, once constructed, can be used as the information source to drive an integration engine. This aspect of MOF is unprecedented and extremely useful.

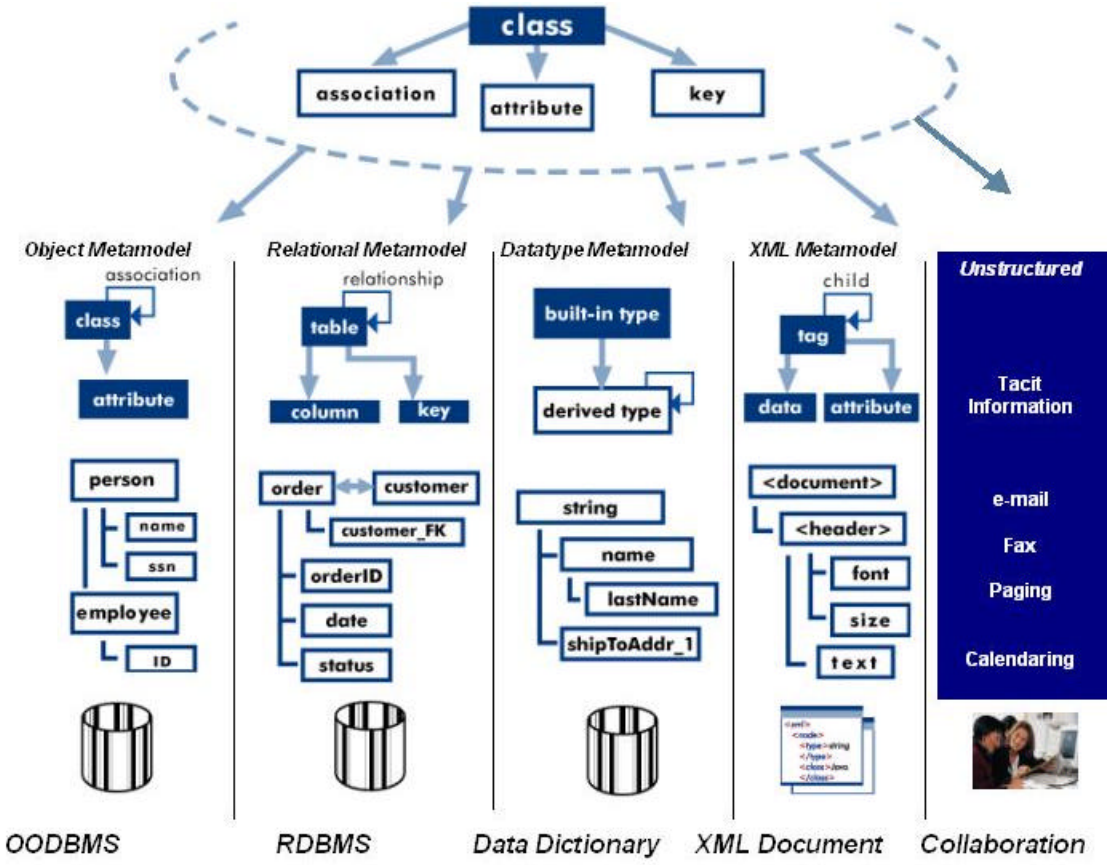


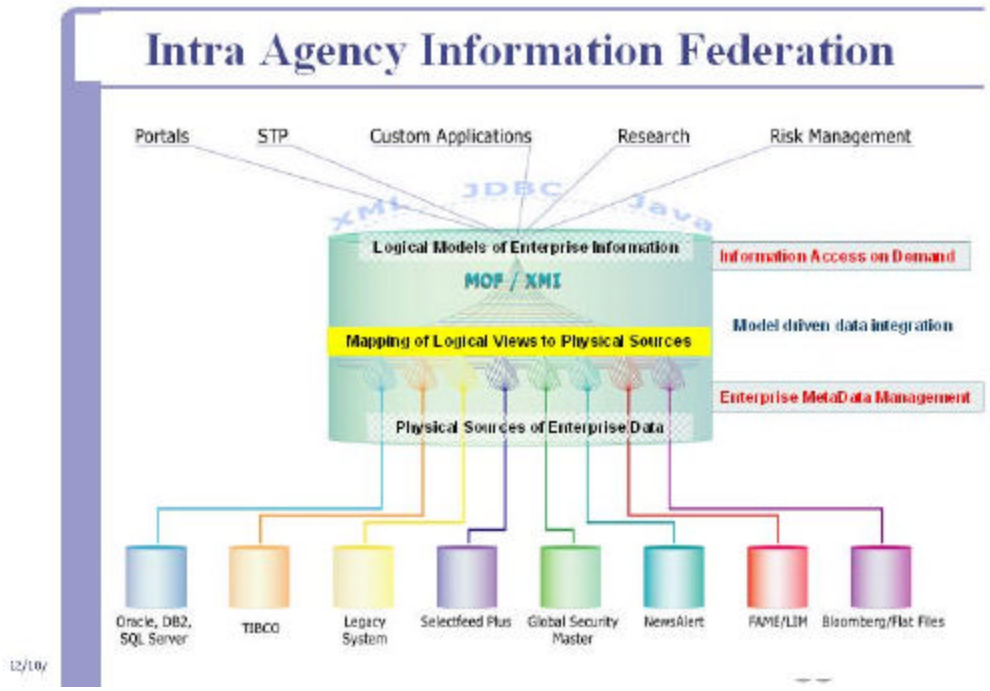
Figure A-4: Target Federated Information Model - Vision Diagram

In the diagram above the top layer is MOF, and we have represented four of the constructs available in the MOF language, one of them is Class. This construct is used to define the construct **Table** in the relational metamodel, the construct **Built in Type** in the Datatype metamodel, the construct **Tag** in the XML metamodel, and the construct **Class** in the object metamodel. Since all of these constructs are derived from a common language, we are able to create and manage relationships between components of the underlying models and to have the "native" constructs appear correctly in the modeling tool.

Any modeling system (whether or not it is MOF-based) deals multiple levels of models. Figure 7 shows the many types of models and levels numbered at the bottom as with M1 and M2. M1 is the system's data and M2 is defined via hard-coded software. This may be why many applications only have one meta-model (for example, Rose and ERWin are based on single meta-models; UML and relational, respectively). Metadata management and modeling

disparate data sources in an easy-to-use manner requires multiple meta-models permitting each model to be built by the user using the language (i.e., meta-model) that best suits what is being modeled.

The key distinction between MOF and non-MOF systems is the following: if there is no way of relating the different meta-models, then the models built with those disparate meta-models cannot be related nor integrated. This is exactly why there have been tremendous difficulties integrating disparate data sources! MOF systems have higher levels of models called M3, which is a language that describes the format and semantics of all of the meta-models. This level of models can be used to identify a "relational table" and to determine that conceptually it is the same as an "object-oriented class" or one of the other data representations such as a "hierarchical record" and an "XML complex type". This technology can be implemented in phases on a layer-by-layer approach. By simply adding additional meta-models, the modeling behavior of a product, and the level of information integration can be expanded very easily.



Appendix B: Services Layers and Capabilities for Information, Data, and Content Management

For the services layer to be successful a layer focused on the pure mechanisms we anticipate will be needed by the Business oriented Services must be made available in the Info Fabric; this we call the delivery layer. This layer provides the standard means of communication among different software applications, running on a variety of platforms and/or frameworks and provides the means to de-couple the services layer from the physical realities of the data it must process. Beyond that the services layer will provide a common set of description, connection and access capabilities for information and data that can be shared by the services that will use it. Information and data will be stored in collections or containers. The resources used and connection with the "Containers" will include a self-describing container model.

Information services can be common and consolidated within two layers we will together call the Info Fabric and separately be designated as the delivery layer and the data layer. The Delivery layer is a platform for the construction of services and is expected to closely resemble the architecture being described in the W3C Web Services Architecture with additional focus on Model-Driven aspects and dynamic data management. The Data layer is a Federated Data Management capability focused on abstracting the nature of the actual physical systems that house and manage data. Together these layers will provide the underpinnings necessary to build the services required to respond to the anticipated Service Reference Model Specific data connection and access services are recommended for addition to the Service Component Reference Model along with how to divide the Information and Data into component elements and address the critical governance issue of ownership and quality responsibility.

The data layer represents information and data as a collection of views, which can be thought of as business objects, XML documents, components, or access and information management services. These views, potentially comprised of information from variety of sources, are used to facilitate the smooth transition from agency centric to Line of Business centric information access and sharing.

To achieve its goals the Info Fabric is an open-standards based framework reliant upon several key standards that have emerged and matured over the past 6 years:

Meta Object Facility - MOF

XML Metadata Interchange – XMI

Extensible Markup Language – XML including related XML security and data access standards such as Wquery, Xpointer, Xlink and new content related standards such as Universal Business Language.

Web Services and Registry Standards – WSDL, UDDI, ebXML

Database and application connectivity standards - JDBC, JCA

Underlying Information and Data Standards- such as SQL for Relational Data Bases, CWM for Data Marts and Data Warehouses, and XML-based formats for semi-structured information such as the Dublin Core, Universal Business Language along with others that be specified by government focused OASIS and OMG Government Task Forces and specific line of business oriented activities.

Notwithstanding this, for the Info Fabric framework to be fully successful it must be designed with standards evolution and innovations in mind. As such it is capable of responding to, and incorporating innovations as they emerge and become hardened for enterprise use.

Appendix C: Security and Privacy with Federated Data Management Approach

An important element of federating of information systems is to have a number of levels of authentication, authorization, and access control that provides a 'defense in depth" approach while providing ease of administration. The security and privacy around a cooperating set of autonomous "components" of information, data or content can include loosely coupled portions or common shared information that has more tightly coupled aspects to it. Access control is very important in this type of federated architecture to the authorization autonomy that is the essential part of federation. With a fully federated approach the federation (business line hub) must authenticate and authorize each access. This can create an extensive amount of overhead. On the other end each federated user has a profile that allows certain subjects and types of access to be done. The user profile of groups of profiles defined by Roles is a critical element of a manageable control approach that can be administered by a few sets of administrators each with separate duties and responsibilities.

One of the key basic premises within a federation is that there cannot be full trust between a federation and the data- information components that they access. Another key capability is to have the ability to handle dynamic needs with the minimum of overhead and human intervention. The system also has to use a set of high level to fine grained access control mechanisms. This is an extensive research area with significant leadership by the Center for Secure Information Systems at George Mason University. The concepts presented here are based on translating their work into the Business Line Architecture and Federated Data Management approach. (A prototype of these concepts is recommended.)

The federated data within a business line may look like Figure C-1. The Federated Schema, also called Virtual Schema will include a set of Authentication, Access Control that will be linked to the related external schema and federated schema. Figure C-2 provides an overview of the access control approach. A key element of our approach is to define a set of Roles within the Business Line and related tasks related to those roles. Work by Pete Epstein and Ravi Sandhu at GMU provides Role- Permission Access Management (RPAM) are especially relevant to the business line –federated data management needs where permissions are mapped to tasks which are then linked to common work patterns, jobs, and roles. The process also separates out support roles for a user administrator, role administration, and system administration along with clearly defining role engineering based on Meta Model for process-oriented Role-Finding. These activities fit with an overall approach of integrating security architecture with enterprise architecture. Figure C-3 provides a brief scenario of how a business line could federate data sources from government agencies and businesses around health data monitoring around a SARS situation with separation of support roles

between Alice, Bob, and Mary. A critical step in the business line architecture and federated data management is that security and privacy are thought of from the beginning during the discovery of the needs and establishing the high level agreements because the security administration must be performed in shared way with clear responsibilities and management of federated user access with personnel joining and leaving and changing their roles and responsibilities.

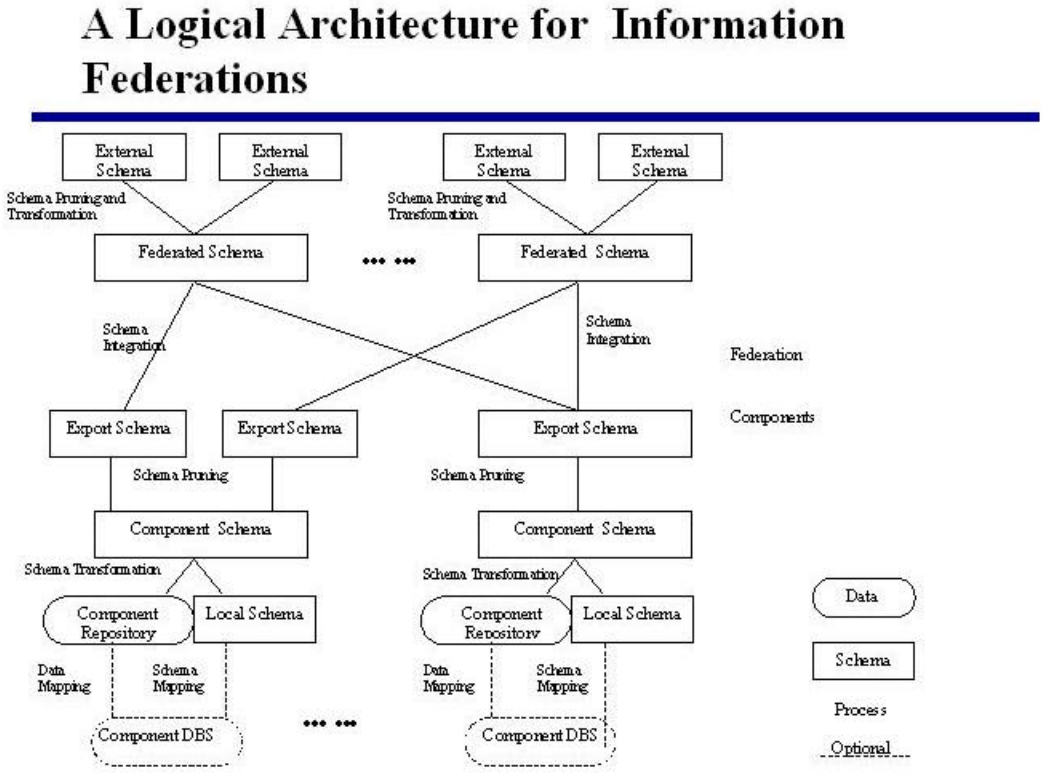


Figure C-1: Logical Model Links Global- Federated Views to Local-Component Views

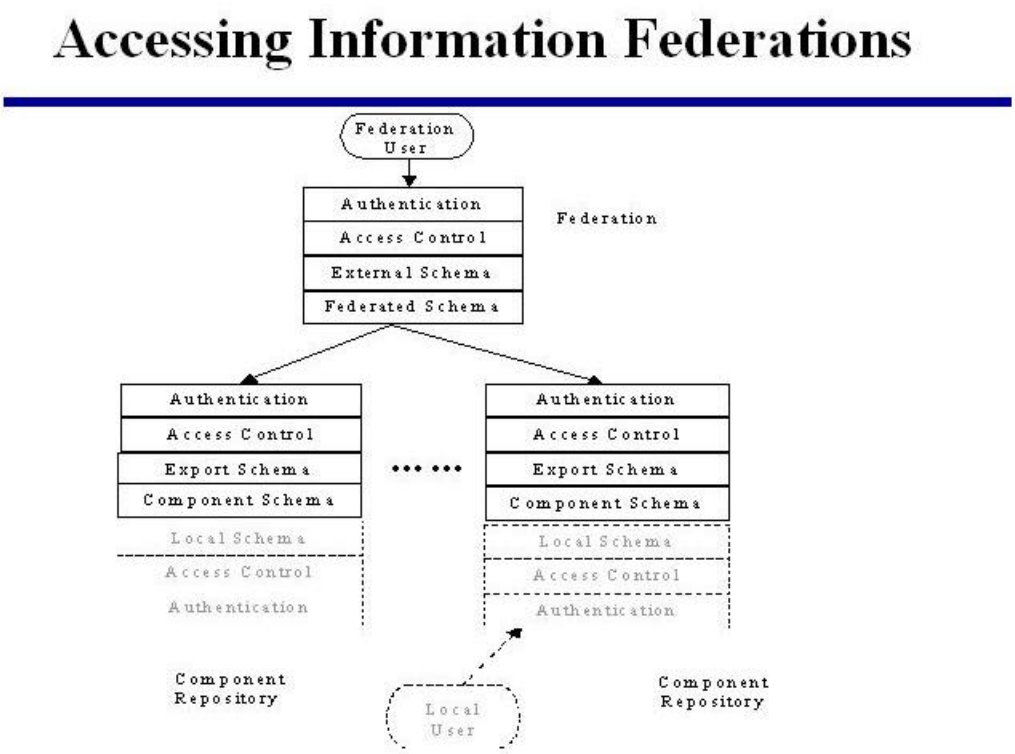


Figure C-2 Multi-Level Access Control Can Provide "Defense in Depth" and handle both Loosely Coupled and Common Shared-Tightly coupled Security

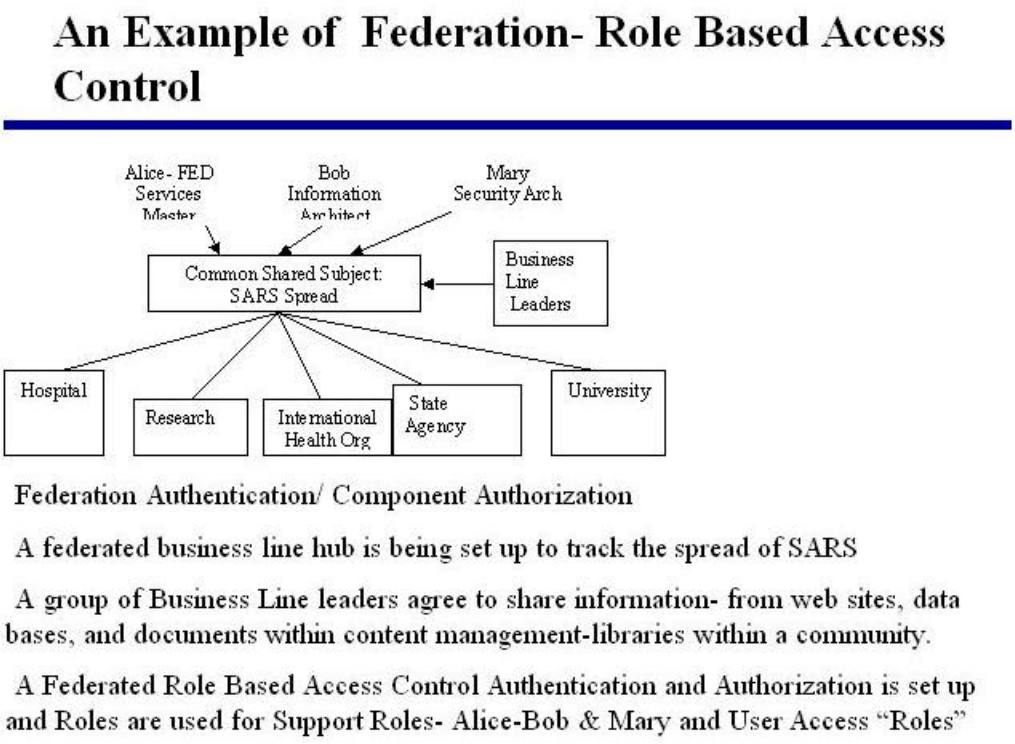


Figure C-3: The combination of Federation with Role Based Access Control must be enabled by Security Administration and Duty Separation.

Appendix D: Information Value Assessment and ROI

D.1 Return on Investment

The data architecture must fit together with an enterprise data management strategy that is driven by a total cost of ownership model where trade-offs can be made from a business centric point of view. This must address not only the cost of organizing and gathering the data, discovering and access data, but also all the maintenance, security and continuity of operations aspects of information, data, and content management. Information, data, and content must be treated as true assets that must be managed and leveraged with the enterprise and as collective resource of the business lines. A technology roadmap and series of alternative paths can be used to reflect the current state of practices, state of the art (best practices) and the emergence of technologies

D.1.1 Lower Cost of Ownership (or as compared to other approaches)

As compared to other techniques that could be used to align the information architecture with the line of business requirements, data federation represents a less costly, more adaptable approach The federated architecture approach does not preclude the use of other key technologies and can include such technologies, as a matter of fact they have been elements that have lead to the approach we are recommending. Our vision is that critical technologies such as Extraction Transformation and Loading (ETL) approach and Enterprise Application Integration (EAI) must be worked into a complete framework based on an overall enterprise data management process and activity models and the types of performance improvements that are needed. Performance results (Customer Results and Mission Results) for a set of changes that improve the data management process and activities. This can result in a set of business cases or sub-business cases that drive a set of improvements that can be shared among the broad group of data related projects. Currently there are many organizations discussing data management issues. Collecting these recommended common issues, defining a baseline process, an information, data and content management life cycle cost and mapping of technologies and values is a key starting point. Technologies can be looked at in a more systematic way with each technology type must be sold on the value it has to support the business needs.

D.2 Increased Efficiency in Application Development

The ROI from a Model Driven Approach to Federated Data Management derives from several factors. Most importantly the models of physical systems are reusable components in all new application development programs. Much of the time spent developing new applications relates to the developer or project manager discovering where the data to support the requirements resides and then understanding the structure of the source systems and their interfaces. Being able

to discover the attributes in models along with the information as to how appropriate they are for a particular use eliminates a lot of time from a development project. The state of the market now is that this sort of activity is repeated every time for each new project. Once the models and repository are created, the activity occurs only once.

D.3 Improved Time to Market

Analyst organizations like Meta Group and Aberdeen estimate that up to 70% of the time required building a new application is related to discovering where the data is to support the application and then writing the code to access and transform it. EII eliminates the bulk of this activity, as it becomes a modeling problem not a coding problem.

D.4 Improved Data Quality - Elimination of redundant copies of data

The hardest part in trying to eliminate redundant systems is recognizing that you actually have redundant systems. Over the past 15 years as data warehousing has become mainstream, data marts have emerged as the data integration standard. Whenever a new application is developed or some new class of report is demanded, a data mart is built to host all of the data required for the mission. It may well be that that all or most of the data has already been assembled in another mart, but there is no way to discover it. Or, the data may exist in two or three existing marts, but again there is no way of finding it. Over time this approach leads to the proliferation of many redundant copies of data being managed, which in turn leads to serious data quality issues as these copies get out of synch.

D.5 Reduction in infrastructure costs

EII can be used to present the developer of a new application the required schema in the form of virtual metadata as a view or document. Instead of building the *n*th data mart, she accesses the virtual object. Redundant copies of data can be discovered, but more importantly, going forward the need for new data marts will be greatly reduced. Building these marts is expensive, but not as expensive as maintaining them. There is tremendous ROI available from the EII approach in reducing the maintenance costs of redundant systems.

Appendix E: Issues and Trade-offs

E.1 Security

Probably the most important issue raised by federation is security. The security of the information is controlled now by each system, which has its own security model. Once federated, the system security model may lose control of the security process. Of course it is still possible to have a particular system still authenticate users and be available through the federated framework. As part of the business line agreements the data ownership and security responsibilities and the role/permission assignments must be made. Role Based Access control concepts based on the work at GMU and the NIST models need to be translated to this environment. A set of the Role Based access document appears in references 22 – 27.

A necessary component of Federated Data Management is a system independent authentication capability. Fortunately there is a standard named X.500 that describes how such a system functions, and there are numerous offerings of such systems available.

Additional information on the Security and Privacy implications of this approach are part of the Security and Privacy Services Framework paper currently under development.

E.2 Connectivity Standards

Two of the most difficult problems encountered in any project to integrate diverse systems are the nature of the connection to the system and then discovering the structure or schema of the system. There are a number of connectivity standards that specify how to connect to and access information (metadata) from systems, these include:

- Java Database Connector – JDBC
- Java Connection Architecture – JCA
- SOAP and related Data bindings
- Web Services

The Business Line Architecture and Integration papers provide additional guidance regarding the use of gateways and Business Line Hubs to address these elements of the solutions.

E.3 Ownership – Stewardship and Deployment Architecture Issues

Once the information becomes federated the issue of who owns it and who is responsible for it must be resolved. The systems that are operated by the agencies

are the originators of the federated data and the agencies must continue to maintain the integrity, quality and operation of the systems. The task of producing a model of a physical system should fall to the entity that operates the system. Once the physical model is produced and registered in a repository, any properly authorized person or group can view it. The repackaging of these physical models into new representations can be performed by anyone, however, this is a highly specialized skill and one or more properly trained groups should manage it in a semi-centralized manner. The agencies themselves should have people skilled at virtual modeling to facilitate the migration and integration of systems internal to each agency.

These semi-centralized groups may provide some or all of the services described below and act as the administration function of a Federated Data Management Center:

- Information and Data Policy – Security, Privacy, and Governance
- Query and Access Control Services
- Data Quality Management Services
- Information and Data Integration Services
- Data Migration: Extraction, Transformation, and Loading
- Information and Data Stewardship
- Data Dictionary and Metadata Management
- Information and Data Administration
- Translation and Mediation with Data Standards and Rationalization
- Information and Data Security and Privacy Services

Figure E-1: Federated Data Management Center Supports Multiple Information Profiles

One possible architecture deployment approach for the federated data management approach is to define a series of federated data management centers that supports the Meta model management and provide the access and data management functions including security and privacy. Other deployment and ownership issues will have to be agreed upon. It is recommended that this be done in an initial pilot workshop and that a project be "designated" to use these concepts building on the experience from the intelligence community, State of Pennsylvania, and the financial service industry.

E.4 Choice Point Management

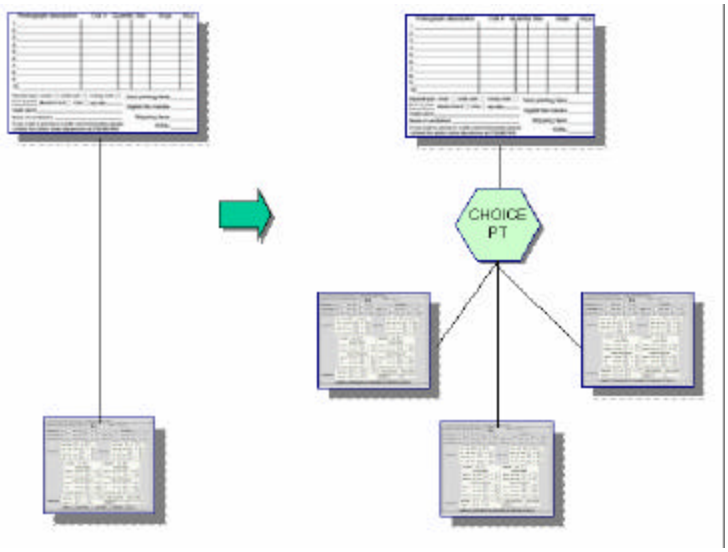
One lesson that has been learned is to offer the uses of the business line community a choice. The Business Line Enterprise environment must foster collaboration rather than mandate if alternatives exist. The danger in a large organization is that parallel worlds and approaches spontaneously exist and do not interact. Studies of our lessons learned reveal that instead of attempting to create homogeneous situations, it is best to recognize, and build mechanisms for working within, heterogeneous environments.

Change drivers are frequently due to shifting customer demands, new operational efficiencies, technology adoption, and consolidation. Because of this, standards are a mixed bag – and a mechanism needs to be placed above them to determine which standard and when used per context of the larger business viewpoint. When vendors seek to gain competitive advantage and market inertia through development of complex standards, productivity can be impeded. In other cases, simple and intuitive standards facilitate rapid deployment.

There are many physical interfaces within an organization, and how these separations work impacts its business functions. Within large organizations, decisions involve thousands of variants of business choices, business rules, business patterns, and data permutations. Organizations need to manage these Choice points in a proactive manner, capturing both options and their rationale. The intentions can then be stored and reused for efficiency and refinement. The explicit identification and management of these Choice points significantly aids to comprehensibility, alignment, while promoting tracing and accountability. In large organizations, the vectors at each decision point, and their interrelated linkage can become quite complex. An agile organization extracts these relationships as business patterns and separates the Choice point vectors out as parameters for each context.

The declarative approach of using templates improves comprehensibility and reduces the probability of errors, as processes are orchestrated based on user's choice. Returning to our Web services question about which Web service gets invoked depends on the business pattern and choices made. The Choice point calls the service and supplies it with the proper parameters. Unfortunately these Choice points aren't typically managed and many times completely ignored in enterprise planning or design. One can only imagine the convoluted deployment of Web services within these organizations; always behind the eight ball as they attempt to control their business rules so that change can occur. On the other hand, an agile organization understands clearly that to avoid change paralysis, managing these Choice points provides a critical baseline and is used for future impact investigation and gap analysis. Today, organizations must develop for choice.

We propose the Government adopt a 'contract' to formalize the combination of workflow, processes, schema, maps, rules, etc. into managed artifacts. The underlying principle is that each layer is to solve its problem, and only its problem, based on a focused set of constraints. Information that is not known during design or runtime is deferred up a layer – thereby providing "help from above". In between layers is a Choice point, providing the information or invoking the appropriate templates based on selected patterns and formalized by the contract.



The specific combination of products and their interrelationships determines the templates needed to generate decision points and variables across an identified pattern. Contract instantiation creates objects at run time that interact as described by the contract; e.g. Web services. Through the use of contracts, dissemination of change from the requirements through to implementation is greatly simplified.

Unfortunately, development tools in this area are relatively immature and will require architecture and design deliverables that sit outside of today's CASE tools. It is anticipated that these tools will begin to incorporate Choice point decision mechanisms in future releases, incorporating capabilities from aspect-oriented development efforts. Until these tools mature in this area organizations will need to improvise and augment their current tools and architectures. The OASIS BCM TC effort underway has been tasked to provide an industry specification on these linkages and specifically the detailed definition of a Choice point service.

Three types of Choice points for both design and runtime are required: (1) *Nouns* or data (BCM artifacts included), (2) *Verbs* or processes and workflows, and (3) *Relationships*, from data models linkage to Contracts themselves. The mechanism can be internal to a product (modeling, mapping, etc.) or as a Web service to the Enterprise as part of a comprehensive infrastructure. The service essentially returns the parameters based on state of all input vectors as defined by the relationships and business rules as declared by the business user.

Experience teaches us that today's organizations are too complex to be modeled with lines and boxes. Current modeling techniques are adequate for showing sub-classing, path options, sets of codelists, or object-role variances; but they fall short in tracing the thread of user choices. This is where the BCM differs significantly from current methodologies as it directly embraces and provides support for choice.

E. 5 Information and Data Quality Management Services

To assure high quality of data within and among the Federal agencies' information systems, a data quality process must provide agencies with a systematic and repeatable process for the detection of faulty data, improving its quality, certifying (statistically measuring) its quality, and continuously monitoring its quality compliance. It is important that the concept of continuous monitoring of data quality be both understood and adhered to for a successful data quality effort. Although reaching required quality levels is a major achievement for a data environment, it should not be construed as the end of data quality effort for that environment. Once the state of data quality is reached, it needs to be maintained. Continuous monitoring is the mechanism by which we can manage the quality of our data with the ever-present possibility of data corruption.

The ultimate outcome of this ongoing systematic effort is the ability to reach and maintain a state in which government agencies can certify the quality level of their data. It will assure the government agencies' data consumers, both within the internal boundaries of the enterprise, as well as those who use the agencies data to function within their own businesses, of the credibility of information on which they base their decision.

In general, four levels of data quality checks should be performed. The following table indicates the four levels, some examples of defect types that can be checked within each level, and the sources from which the quality assessment logic can be generated. Each level can include several characteristics (dimensions) of data quality. Characteristics may vary depending on the type of data store being analyzed, its purpose, and its specific user quality expectations.

ANALYSIS LEVEL	DEFECT TYPES	CRITERIA
Level 1 - Completeness and Validity	• Completeness - The degree to which values are present in the attributes that require them. • Validity - The quality of the maintained data is rigorous enough to satisfy the acceptance requirements of the classification criteria; A condition where the data values pass all edits for acceptability, producing desired results.	Data environment descriptions
Level 2 - Structural Integrity	• primary key / defining attribute integrity - The ability to establish the uniqueness of a data record (and data key values) • referential integrity • cardinality	Data environment data model (explicit or implied)
Level 3 – Business Rules	3 rule violations	Rule repository or

ANALYSIS LEVEL	DEFECT TYPES	CRITERIA
		business experts
Level 4 - Conversion Rules	4 rule violations	Data mapping specifications

E.5 Information Profile

Information Profile encompasses the following five properties:

- Information Type: There are three major types of information: Structured (Tabular); Semi-Structure (XML documents); Unstructured (Tacit, Collaborative, email)
- Usage: Information can be used either as the transactional or analytical or reference
- Distribution: Information can be distributed in three ways: Centralized; Distributed; Hybrid
- Storage: Information storage includes relational; object-oriented; cube; sequential; indexed sequential; Hierarchical; and network
- Location: Government's information could be located at the Agency; on Internet; in any Government's organization/department
- Services based on the proposed Security and Privacy Framework will also be included

Federal Government's Information at Sub-Function level can be collected and stored in a repository based on the Information Profile defined in this section. This way the government's information will be collected uniformly at one time and will be distributed in desirable ways to the citizens, governments units, and business partners. The Information Profile structure is depicted in Figure 10.

E.6 Unstructured Data

Much of the information stored in IT systems is classed as unstructured, generally pure text. There are methods available for "integrating" this sort of information with data from structured sources. Repositories of unstructured text can be tagged by COTS products designed for this purpose, or by humans who are expert in the content categories of the documents. In either case, once tagged, these repositories can be "modeled" and then integrated. This tagging usually involves more human effort than is justified to make a repository of unstructured content "joinable" with structured content.

Commercial search engines can be used to "integrate" vast collections of unstructured content (the World Wide Web) by implementing proximity search technology across strings of characters, but this does not solve the problem of integrating this sort of content with databases and file systems. We recommend that the need to do real integration of structured and unstructured be examined on a case-by-case basis, and when justified, apply the automated tagging technology required to impart "structure" to the text.

Appendix F: Referenced Documents

1) *The Business Reference Model, Version 1.0* February 2002, Federal Enterprise Architecture Program Management Office
2) Busse, Kutsche, Lesser, Weber, Federated Information Systems:Concepts, Terminology and Architecture, 1999
3) Federated Information Systems - Conferences, 1999, 2000, 2002
4) OMG- Meta Object Facility, UML 2.0, MDA, XML Metadata Interchange (XMI) www.omg.org
5) Intelligent Information Integration Study- DARPA- 1995
6) Amit Sheth and Larson: Federated Information Systems, 1990. University of Georgia.
7) Larry Kershberg, and Amihai Motra: VirtuE: Virtual Enterprise Environment, George Mason University.
8) Professional ebXML Foundations; Bruce Peat, et al; 2001 WROX Press
9) OASIS ebXML Registry Standard, Katherine Breininger, 20 Jan 2003
10) Ref: Technical Overview of OASIS/ebXML Registry v3.0, Farrukh Najmi
11) Ref: XML Collaborator: XML Design Collaboration And Registry Software White Paper September 2002, Kevin Williams
12) DOD Net-Centric Data Management Strategy, V14, March 4, 2003.
13) OMB's Role in Overseeing Information Quality, John D. Graham, March 21, 2002.
14) Agency Draft Information Quality Guidelines, John D. Graham, June 10, 2002.
15) Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies, Sept. 5, 2002.
16) A XML Schema for Electronic Record Management, Leila Naslavsky, Dorrit Gordon, University of California, Santa Cruz, 2002.
17) Continuing Criminal History Records Improvement Evaluation, Final 1994-1998 Report, February 2000.
18) FAA Data Improvement Initiative, 2002.
19) Justice XML Data Dictionary, Schema, and Registry/Repository: Overview and Status, John Wandelt, Mark Kindl, Jan. 14, 2003.
20) Public Health Data Model, Frequently Asked Questions. July 2000.
21) Crossing the Structure Chasm, Alon Halevy, et. al, Jan 2003.
22) Recommended XML Namespace for Government Organizations, Jessica Grace and Mark Crawford, March 2003.
23) Role-Based Access Control on the Web, Joon Park, et. al, ACM Transactions on Information and System Security, February 2001.
24) Engineering of Role/Permission Assignments, Pete Epstein and Ravi Sandhu, 2000.
25) Framework for Role-Based Delegation Models, Ezedin Barka and Ravi Sandhu, 2001.
26) Engineering Authority and Trust in Cyberspace: The OM-AM and RBAC Way, Ravi Sandhu, 1999.
27) Towards a UML Based Approach to Role Engineering, Pete Epstein and Ravi Sandhu, 2001.
28) The NIST Model for Role-Based Access Control: Towards a Unified Standards; Ravi Sandhu, David Ferraiolo, Richard Kuhn, 1999.
29) Enterprise Transformation – Agile Solutions Requires Developing for Choice; Audrey Davis, DFAS CIO, 2003 http://www.dfas.info
30) Berthold Daum, Modeling Business Objects with XML Schema, 2003.
31) IAC Business Line Architecture & Implementation White Paper March 2003.