



August 8, 2008

# Enabling Information Discovery and Access

---

*Concepts for Context Metadata Management within the Federal Community*

- Bryan Aucoin  
President, Alanthus Associates, Inc.  
[bryan.aucoin@alanthus.net](mailto:bryan.aucoin@alanthus.net)  
703.684.9888 (o)  
703.307.3191 (c)

## Introduction

### The Purpose of this Paper

There is an emerging recognition within the Federal Community, along with a growing body of legislation, regulation and policy guidance, that organizations need generalized technical approaches to enable information sharing.

Posting information to web sites to permit discovery by search engines is a place to start. However, search engines by themselves are not enough to fully enable data access and discovery. Access to sensitive data must be managed, generally through the systems that are used to create and manage them, and therefore the data cannot be directly ingested into search engines. Hence, it becomes critically important to make data context (i.e. the metadata, the data about the data) discoverable and accessible within any broad, federated enterprise.

This paper describes one generalized approach for discovery and access of data context in accordance with the legislation and policy guidance.

### Knowing What we Know

Irrespective of the amount of time, labor and resources devoted to information sharing since 9/11, it remains an intractable problem within the Federal Community. While we have made enormous progress, our ability to know what we know and then utilize what we know remains elusive<sup>1</sup>. There are a wide range of best practices, but there are few consistent, generalized approaches that have been adopted across the Federal Community. However, there is a growing recognition within this Community of the need for such generalized approaches. Consider the following extracts from various Federal Community policy documents:

---

<sup>1</sup> At the Federal-level, information sharing is generally viewed within the context of “federated enterprises”. These federated enterprises include our Federal Government; the Department of the Defense and our National Security Community; State, Local and Tribal government entities; international and coalition partners; and public, private, and academic sector organizations.

*“The DoD will leverage the many investments it has made, and will continue to make, within the net-centric and the information technology (IT) environment. The DoD will work with external partners who have highly developed technology support for information sharing while maintaining capabilities to work with external partners who have limited or no technology support for information sharing.”*

- DoD Information Sharing Strategy  
May 2007

*“The Intelligence Community collects a vast amount of intelligence information from many sources, information that is difficult to discover or access outside of collection stovepipes. Analysts “don’t know what they don’t know.” They are often unaware that information has been collected. We need to move to a collaborative information environment, where all information is discoverable by Intelligence Community collectors and analysts, relationships between information can be easily discerned, and the people and organizations of an integrated enterprise readily engage one another to synthesize knowledge.”*

- IC Information Sharing Strategy  
February 22, 2008

*“Data Sharing is a standardization area within the DRM. A Community of Interest (COI) should have common capabilities to enable information to be accessed and exchanged. Hence the DRM provides guidance for the types of services that should be provisioned within a COI to enable this information sharing.*

- Federal Enterprise Architecture Data  
Reference Model  
November 2005

In short, there is an emerging consensus of opinion, and the time is ripe to forward the debate about common technical approaches to information sharing.

## Search Engines and the Limits to Information Discovery and Access

One primary focus within the Federal Community has been making data discoverable by and accessible to the citizens of the nation<sup>2</sup>, and search engines play a prominent role in these efforts. However, search engines have certain limitations. Simply and obviously, they can only make the information that they can access available for discovery. Search engines can only make discoverable what they can “crawl”; they can only ingest those web sites that they can readily access and index.

Some have argued for a seemingly logical next step: “Let’s just make all the data available for discovery by search engines.” And, the immediate counter is that as citizens of a democracy, we do not want all the information that our government holds about us to be readily discoverable. Publication of individual tax information would make us targets for identity theft. Publication of our medical information would violate our privacy and may make many subject to prejudicial treatment. Disclosure of proprietary company information would compromise our ability to compete in an international economy.

So, the question then becomes: “How do we strike the right balance?” How do we make enough discoverable so that we can share effectively, efficiently, and without compromising critical data?

## An Architecture for Information Sharing

### Knowing That We Know:

### Understanding the DRM Concept of Data Context

The Federal Enterprise Architecture Data Reference Model (FEA DRM) introduces the concept of “data context”, in part, to address these types of issues with data discovery and access.

The FEA DRM uses an accurate, albeit somewhat esoteric definition for data context. Specifically:

*“The Data Context standardization area facilitates discovery of data through an approach to the categorization of data according to taxonomies, and provide linkages to the other FEA reference models.”*

---

<sup>2</sup> A number of OMB memoranda have promulgated and the then reinforced this policy direction:

OMB Memorandum M-05-04: Policies for Federal Agency Public Websites dated December 17, 2004

OMB Memorandum M-06-02: Improving Public Access to Dissemination of Government Information and Using the Federal Enterprise Architecture Data Reference Model dated December 16, 2005

OMB Memorandum M-07-20: FY 2007 E-Government Act Reporting Requirements

The FEA DRM goes on to define “Context Awareness Services” as follows:

*“A context awareness service allows the users of a collection to rapidly identify the context (as defined above) of the data assets managed by the Community of Interest. Context information may be captured in a formalized data architecture, a metadata registry or a separate database.”*

Again, the language is esoteric, intended for the Federal Community of Data Architects. However, the underlying concepts are simple. Data context is metadata (data about data). It is the set of data that information providers should make available for discovery by the prospective information consumers that describes the information that the providers hold. - “Here’s what we know.” (These are the provider’s *Data Assets* in DRM terminology). In developing an information sharing strategy, an organization should determine what information consumers should know about their data holdings and provide a way for them to get to that data<sup>3</sup>. The FEA DRM provides the following guidance on what context data to make available for discovery:

*“Typical examples of Data Context for a given Data Asset may include a Topic identifying a subject area, a data stewardship assignment, sources of record, etc. At a minimum, the Data Context for a given Data Asset should answer the following questions:*

- *What are the data (subject Areas/Topics and entities of interest) contained within the Data Asset?*
- *What organization is responsible for maintaining the Data Asset?*
- *What is the linkage to the Federal Enterprise Architecture Business Reference Model?”*
- *What services are available to access the Data Asset? (See Data Sharing)”*

Again, the basic concepts are simple. Information consumers need to be able to discover what types of information an enterprise holds, who maintains it, how does it relate the Federal Government’s business, and how they can get access to the information once they determine that they need it. In slightly more technical terms, we can separate the context metadata, and make these metadata accessible and discoverable to any prospective information consumer in a consistent, generalized way.

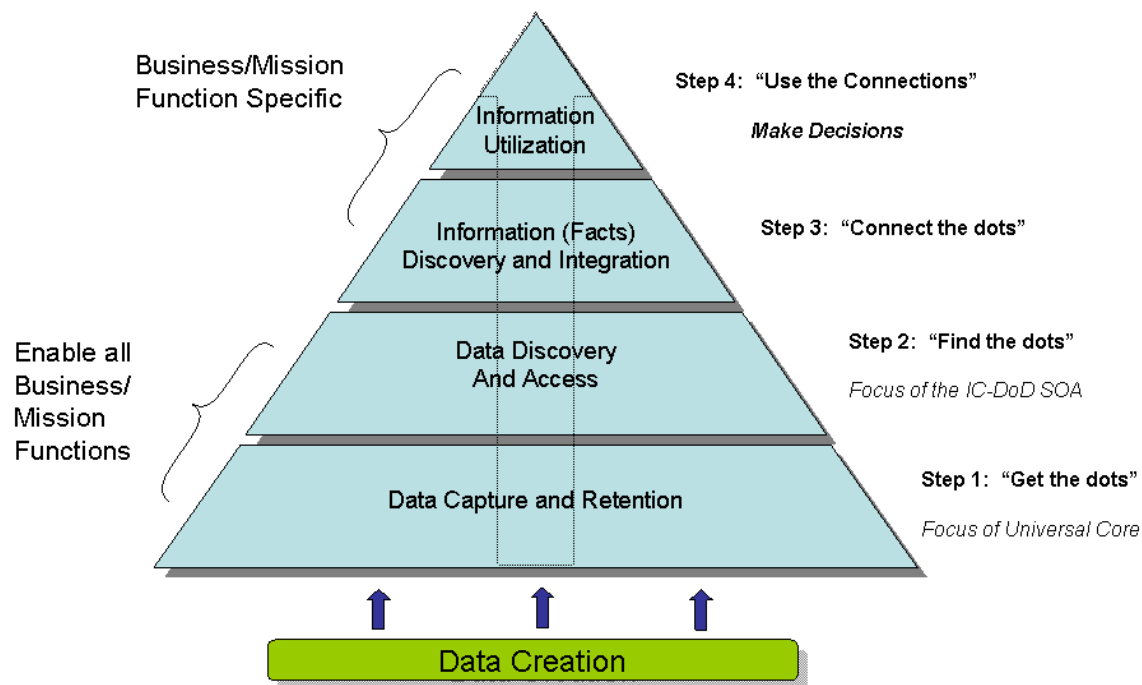
---

<sup>3</sup> Note the Intelligence Community doctrine of “Responsibility to Provide” within the Intelligence Community’s Information Sharing Strategy.

## A Different Perspective: Information Consumers and Information Needs

The preceding section discussed the role of information providers and the actions that they should take to let their knowledge be known. Information consumers have the complementary perspective – “What data do I need to know to do my job, where are they at, and how do I get access to them?” They ask the FEA DRM context questions delineated above.

One way to look at from the information consumer perspective is as a “Data Hierarchy of Needs”.<sup>4</sup>



In general, any information consumer will focus on discovering and accessing the data that they require to do their work; they will integrate the data that they discover to create new information, and then they will take action based upon what they learn.

Enterprises, acting as information consumers, will typically build applications that support their respective business processes. To implement these applications they generally build new backend data stores that are used to capture the new information that they create and/or make the data they use within their business processes readily accessible. In the process, they become information producers as well. They become responsible for making the information that they create accessible and discoverable.

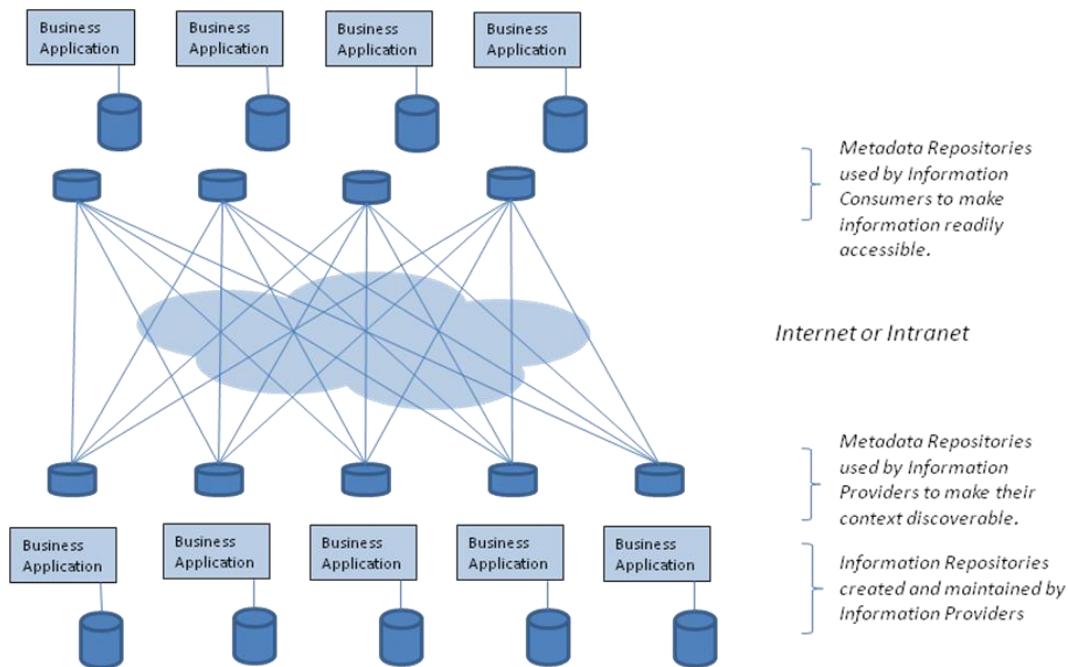
<sup>44</sup> See *Federated Governance of Information Sharing Within the Extended Enterprise* dated 15 January 2008 by the Association for Enterprise Integration Information Sharing Working Group. <http://www.afei.org/documents/ISWGWhitePaperfinal.pdf>

As an example, the organizations responsible for records management within a Federal agency will create and maintain information about the official correspondence that that agency produces, including record retention schedules and responses to Freedom of Information Act (FOIA) requests. They may retain reference copies of official correspondence to facilitate their access. In turn, they may also make these types of information available to the broader enterprise. These organizations will find the official correspondence, they will create and manage metadata about these documents to support their business processes, and they will generally provide this information to broader enterprise.

### An Emerging View: Toward a Consistent, Generalized Technical Approach

Looking at industry trends, we can pose what a generalized technical approach will serve the needs of both information consumers and information providers. Metadata repositories will serve a central role in both provisioning the context information and in making commonly referenced data accessible for reuse.

A high-level generalized architecture is depicted below:



Metadata repositories provide an efficient means for information providers to permit the context information for their data holdings to be made discoverable and accessible. Further, these repositories provide the means for information consumers to capture and organize the data that they use within their business processes in a simple, convenient way. Metadata repositories become complementary to search engines. - Search engines can ingest the context metadata for a given set of information holdings

to make the subject matter held within the holdings discoverable, as well as the policy and services used to access them.

The FEA DRM poses a range of options for providing context awareness (data architectures, metadata registries, databases) by intent. Best practices had yet to be determined. However, XML and XML server platforms provide an efficient solution. Context metadata tend to be “document oriented” (using the FEA DRM language). With XML, the data are stored in the same format as used for exchange, meaning that no mediation is required when moving the data around. This simplifies the implementation of context awareness (web) services. Finally, because of the document orientation, the data can be readily ingested by search engines.

## **Conclusion: Use of Metadata Repositories:**

Metadata repositories, and specifically XML servers, can be used both as a means to make context information discoverable and accessible and as a means to make commonly referenced information readily accessible. For a Federal organization, it simplifies compliance with the OMB directives and provides a simple, direct means to implement FEA DRM guidance. It enables search engines to find relevant context information and the policy and points of contact for gaining access.

For any large, federated enterprise or a community of interest, it makes it easy for information producers to make their data discoverable. It provides the means to “know that we know” and a ready means for information consumers to determine how to gain access to the information once they have discovered that it exists.

Once an information consumer has gained access to the data they need, then they can use metadata repositories to capture the new metadata that they create as part of their business processes.

*The development of this paper was sponsored by Mark Logic Corporation as an independent assessment of the application of metadata repositories for information sharing within the Federal Community. Mark Logic has implemented metadata repositories to support both information providers and information consumers.*