Enterprise Data Architecture Strategic Assessment Architecture and Roadmap

Table of Contents

About This Document
Executive Summary
   Figure 1 - High Level Near Term Implementation Plan
Introduction
   Background
   Project Overview and Participants
      EWSolutions
Existing Conditions and Observations
Enterprise Architecture
Enterprise Data Architecture (EDA) / Enterprise Information Management (EIM)
   Data Governance
   Figure 2 - Data Standards Governance
   Figure 3 - Data in Relation to Other Key Assets
   Data Stewardship
   Data Quality
   Data Sharing and Collection
   Metadata Management
   Table 1 - Enterprise Metadata Systems/Registries/Repositories
Data Modeling
   Table 2 - DCM Example Section
   Data Warehousing / Business Intelligence
   Structured Data Management
   Unstructured Data Management
   Master/Reference Data Management
Key EPA EDA/EIM Strengths
Key EPA EDA/EIM Opportunities for Improvement
Recommendations
Overview

Figure 4 - DAMA DMBOK Functional Framework
Figure 5 - EIM Functional Framework

Functional Recommendations

Data Governance

Figure 6 - Example Data Governance Structure

Data Stewardship

Figure 7 - Example of Cross-cutting Data Stewardship

Enterprise Data Model

Figure 8 - EDM levels

Figure 9 - Example Subject Area Model (for Healthcare)

Figure 10 - Example Enterprise Conceptual Data Model (ECDM)

Figure 11 - Example Enterprise Logical Data Model

Enterprise Data Warehouse

Figure 12 - Example EDW

Metadata Management

Figure 13 - Example MME Architecture

Figure 14 - Horizontal/Vertical Data Lineage

Figure 15 - Example Prioritization of Types of Metadata

Data Quality Management

Data Sharing

Master/Reference Data Management

Target Enterprise Data Architecture Document Recommendations

Figure 16 - EPA DRM (partial)

Key Next Steps

Initial priorities

Near-term plan for EDA/EIM initiatives

Figure 17 - Near-term EDA/EIM Plan/Priorities

Key Dependencies

Appendix A: Glossary of Terms

About This Document

This document is a deliverable of the Enterprise Data Architecture Strategic Assessment project, performed for the Environmental Protection Agency (EPA) by Enterprise Warehouse Solutions, Inc (EWSolutions). This project was executed from November, 2009 through early February, 2010.

This high level assessment provides an Independent Validation and Verification (IV&V) for the draft “EPA Enterprise Architecture - Target Data Architecture” (Target Data Architecture) document created prior to this engagement.
While the primary audience for this document is EPA IT management and staff, business executives will benefit from reading many sections of the document. Some material within the document is moderately technical.

Executive Summary

Interviews with EPA business and IT leaders, along with analysis of the draft “EPA Target Data Architecture” document (Target Data Architecture) and system documentation (at a high-level), led EWSolutions to the following conclusions:

1. Enterprise Architecture has been instituted in the EPA based on recommendations of the OMB, with guidance from the Federal Enterprise Architecture (FEA). One of the key architectural components within the EPA EA is the Data Architecture. Formalizing an EPA Enterprise Data Architecture means that a more holistic approach to data architecture is taken which takes the needs of the agency into account rather than an ad-hoc, largely program office-driven approach to data architecture.

2. The draft Target Data Architecture document takes into account the main components of Data Architecture and Enterprise Information Management (EIM). In this context, Data Architecture can be thought of as the vision and strategy for more effective and efficient capture, storage, reuse, analysis, sharing, and publishing of EPA data and information assets. Enterprise Information Management (EIM) includes the activities to realize, implement, and sustain the Target Data Architecture vision and strategy.

3. The Target Data Architecture document is, in general, based on industry best practices and draws inspiration from the DAMA-DMBOK (Data Management Body of Knowledge), of which EWSolutions was a key contributor.

4. In this document, EWSolutions will provide our assessment of the current state, key strengths, areas of improvement, and specific recommendations. Key recommendations, which largely coincide with the Target Data Architecture document, are as follows:
• Institute Data Governance and Data Stewardship to provide the enterprise oversight of data and information (critical EPA resources).

• Create an Enterprise Data Model (EDM) in order to obtain an enterprise view of data resources to improve business / IT alignment, improve reuse, and improve data sharing.

• Create an Enterprise Data Warehouse (EDW) to integrate key data resources for improved publishing, analysis, and sharing of data in order to provide Business Intelligence capabilities not currently or easily possible.

• Formalize a Managed Metadata Environment (MME) strategy. Investigate utilizing existing RCS environment as the foundation for an enterprise MME – taking into account that such a MME will need to be able to integrate (physically or virtually) a wide variety of metadata sources.

• Capture, store, and publish Data Quality statistics, especially regarding external data providers.

• Investigate using direct database connectivity and Enterprise Information Integration (EII) software as alternatives for improved internal data sharing.

• Continue and expand development of Master/Reference data stores

5. EWSolutions recommends that the EPA Enterprise Data Architect makes refinements to the draft Target Data Architecture based on the above (and more specific recommendations below) and submits this to the appropriate governance body for approval.

Below is the proposed near term (less than 2 years) implementation plan to realize the Enterprise Data Architecture after the Target Data Architecture is refined, approved, and promoted:

Figure 1 - High Level Near Term Implementation Plan

Below is a short list of the critical initial activities/projects for the implementation of the Target Data Architecture:
• Refine, obtain approvals for, and promote the EPA Target Enterprise Data Architecture document

• Develop DAC (data governance) Charter, identify Executive Sponsor, Chief Business Data Steward, Executive Stewards (representing an AA'ship, regions), obtain approvals for and promote the DAC Charter.

  • Develop an Enterprise Data Modeling (EDM) strategy

    • Review/refine/approve the existing Data Classification Reference Model (DCM)

    • Obtain approvals from DAC & QIC for the EDM strategy and DCM

    • Coordinating Data Stewards identified by DAC based on DCM subject areas

  • Develop an Enterprise Data Warehouse strategy

  • Form Data Stewardship Committee; identify data stewardship policies and tools

  • Develop a long term MME strategy, confirming whether RCS to be the foundation of the MME (Managed Metadata Environment)

Introduction

Background

The Environmental Protection Agency (EPA) mission is to “protect human health and the environment”. To help accomplish this mission, the EPA collects, shares, analyses, and publishes data about environmental threats received from a variety of sources, often from external parties such as states, tribes, and industry.

The EPA has a multitude of regulations and directives which it needs to respond effectively to. Recently the EPA has performed significant work in response to ARRA (American Recovery and Reinvestment Act), and is working to address global climate change through it greenhouse gas (GHG) initiatives.

The EPA has instituted the practice of Enterprise Architecture based on the Federal Enterprise Architecture (FEA) and the Federal Segment Architecture Methodology (FSAM). The EPA has been logically divided into multiple, cross-cutting (spanning offices) segments in order to obtain an enterprise view into component architectures - strategic, performance, business, technical, services, and data. The draft Target Data Architecture has been formed in order to realize the Data Architecture component of the EPA’s Enterprise Architecture.

In order to move the Target Data Architecture forward, EWSolutions has been selected to perform a high level assessment which provides an Independent Validation and Verification (IV&V) of the draft Target Data Architecture.

The assessment approach followed was:
1) Develop questionnaire aimed to understand how the EPA’s information environment is aligned to business requirements. There were questions regarding both the current state and future state environment. Questions had a business emphasis, but were tailored to uncover information regarding all the major areas of Enterprise Data Architecture (EDA) and Enterprise Information Management (EIM).

2) Interviews were conducted. Due to the short duration of the assessment, only the following EPA personnel (plus 1 contractor) were interviewed (in group or individual interviews).

LIST OMITTED TO PROTECT CONFIDENTIALLY

3) The questionnaire was distributed to interviewees in advance, and was also distributed to additional EPA personnel who were not able to be interviewed. The following responded in writing to the questionnaire.

LIST OMITTED TO PROTECT CONFIDENTIALLY

4) Documentation review was performed at a high level. Documentation sources included forwarded material, publically available material from the EPA website, and material from the EPA Enterprise Architecture portal.

5) Final deliverables which were produce include this document, a current state assessment presentation, and a target state assessment presentation.

6) A final presentation of findings to be conducted February 2, 2010 at EPA Headquarters.

The following EDA/EIM areas were investigated at a high level in this strategic assessment.

- Data Governance and Data Stewardship
  - Data Sharing and Collection
    - Data Modeling
  - Data Warehousing / Business Intelligence
    - Metadata Management
      - Data Quality
    - Master / Reference Data Management
  - Structured and Unstructured Data Management
    - Information Security

Note: in this document the Enterprise Data Architecture (EDA) will refer to vision and strategy, whereas Enterprise Information Management (EIM) refers to the implementation of the EDA and ongoing implementation management.
Project Overview and Participants

The EPA Enterprise Data Architecture Strategic Assessment was launched November 4, 2009. The project was completed in roughly 3 months elapsed time (taking holidays and vacation and other factors into account), concluding with the delivery of this report and related deliverables.

Kevin Kirby, EPA Chief Data Architect, was the project’s executive sponsor.

Pete Stiglich, Principal Consultant/Trainer at EWSolutions, served as project manager and analyst for the project.

Note: See article I was interviewed for and commented on as well as for this document.

EWSolutions Managing Partner Mike Jennings provided engagement oversight for the project.

EPA Segment Architects, especially Vince Allen, Eugene Durman, Nathan Wilkes, and Lisa Jenkins provided excellent support to the project and helped to arrange interviews and questionnaire responses.

EWSolutions

Enterprise Warehouse Solutions, Inc. (EWSolutions) is a leading consulting firm in the field of Enterprise Information Management. EWSolutions provides strategic consulting, systems integration, and training services with a focus on data architecture, data modeling, data warehousing and business intelligence, metadata management, master and reference data management, information quality, and data governance and stewardship.

Existing Conditions and Observations

Enterprise Architecture

The EPA, with guidance from the OMB and elsewhere, has instituted the practice of Enterprise Architecture (EA) and realizes that Data Architecture is one of the key component architectures in the Federal Enterprise Architecture (FEA) paradigm.

“The vision of this target” (EPA EA) “is to help the Environmental Protection Agency (EPA) move towards a more service oriented architecture, with the intent of enabling more agile application development, leveraging reusable components (services and data), and making our Agency-wide tools and services more interoperable, at lower cost, and with shorter development and deployment life cycles. 1”. (1 Source: FY2009 Enterprise Target Architecture 3.0)

The EPA Enterprise Architecture Working Group (EAWG) has been formed and has extensive, though not universal, participation of program offices. Some have the perception that EA is additional overhead and that the only reason for developing an EA is due to OMB mandate.

The EPA does not currently have an approved and promoted EPA Enterprise Data Architecture which is a critical component of an EA, though a draft Target Data Architecture has been crafted and is the subject of this IV&V. Another way of thinking about Enterprise Data Architecture is as a formalized vision and strategy for collecting, managing,
securing, analyzing, sharing, and publishing data and metadata in a holistic, enterprise fashion in order to serve EPA stakeholders effectively and efficiently.

Formalizing and implementing an Enterprise Data Architecture will provide the EPA with improved opportunities to serve stakeholders through:

- increased opportunities due to more rapid and effective interoperability, integration and data sharing
- reduced costs due to reuse (data, metadata, systems)
- improved business/IT alignment, knowledge retention, communications, and semantic reconciliation
- overall, increased effectiveness and efficiency.

Enterprise Data Architecture (EDA) / Enterprise Information Management (EIM)

The current state of the EPA Enterprise Data Architecture and EIM practices is described below based on each key component.

Data Governance

“Data governance is the practice of making enterprise-wide decisions regarding an organization’s information holdings”

Source: Data Management Study, Federal Enterprise Architecture Program Management Office

There are currently a number of organizations which provides governance and leadership on certain aspects of EDA/EIM – yet, there isn’t currently a central agency-level governance body which has a holistic view or mandate to govern data and metadata, in a manner which spans all program offices, regions, and applications for all types of data (environmental, administrative, financial, HR, etc).

The Quality and Information Council (QIC)’s main emphasis is on governing the Quality System, which is focused on environmental data and technology (but does not seem to cover administrative, financial, HR, or other non-environmental specific data). The Exchange Network Leadership Council (ENLC) is focused on the NEIEN (National Environmental Information Exchange Network) “aka Exchange Network” which deals primarily with data exchange with states and other external entities.

There are also some office specific data governance bodies (in place or proposed), e.g. ORD’s Scientific Data Management Oversight, OW’s Information Steering Committee. Below is an organization chart for Data Standards governance. In the EPA context, Data Standards deal primarily with data exchange between the EPA and its partners.
Data and information are key non-fungible assets of any large enterprise in the 21st century, and yet are often treated as a byproduct of an application. Much data in the EPA needs to be shared and reused apart from the source application – having Data Governance and Stewardship in place can help to ensure that data can be more effectively used. For example, Data Stewards, Data Architects, Data Models, Application Developers, and others can leverage an Enterprise Data Model in order to understand essential data entities and data elements (attributes), so they can build new or adjust existing information systems which are more aligned with the business and more interoperable via the data layer.

Figure 3 below identifies the key assets of any modern enterprise and contrasts how these different asset types should be governed, modeled, and managed.

**Figure 3 - Data in Relation to Other Key Assets**

<table>
<thead>
<tr>
<th>People</th>
<th>ASSETS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Governance</td>
<td>People, Money, Facilities and Equipment</td>
</tr>
<tr>
<td>Model</td>
<td>People, Money, Facilities and Equipment</td>
</tr>
<tr>
<td>Rules</td>
<td>People, Money, Facilities and Equipment</td>
</tr>
<tr>
<td>Tracking</td>
<td>People, Money, Facilities and Equipment</td>
</tr>
<tr>
<td>Security</td>
<td>People, Money, Facilities and Equipment</td>
</tr>
</tbody>
</table>

Adapted from "Building and Growing a Successful IQ Function" IDQ Conference 2007. C. Lwanga Yonke

A common complaint expressed in the interview process and questionnaire responses was regarding a lack of engagement upfront between program offices and the Office of Environmental Information (OEI). Data Governance should be directed primarily by business representatives in order to obtain better alignment with business requirements.
Some quotes regarding the lack of engagement between business and IT include:

“QIC is broken as there isn’t enough engagement upfront with the business when initiatives are coming. Risk identification is important – going to programs with ideas and understanding risks. Risks are typically buried and not talked about in a strategic fashion.” - Segment Architect

“More transparency & more inclusiveness in OEI decision making.” - questionnaire response regarding future state vision

“..orient OEI to respond to program needs in sustainable, customer-oriented approach” - Questionnaire respondent

“A QIC decision might be made a month or two before the Data Architects are made aware of it – decisions seem to stop at the senior management level.” - Segment Architect

Interviewees expressed (directly and indirectly) strong needs for an enterprise Data Governance body to address prioritization, funding, and increasing cooperation and communication.

“..have to have leadership - otherwise OW will solve its problems, OECA will solve its problems, OAR will solve its problems and they won’t talk to each other. The issues cross cut with the other program offices.” - Branch Chief (acting)

Critical Success Factors “(1) Higher and more predictable levels of funding for systems development, (2) more effective communication between IT staffs and program management, and (3) improved OEI initiative and leadership in managing and funding central data resources (registries, enterprise services and tools, enterprise licenses, and IT development and portfolio management tools). “ - Questionnaire respondent

“Even though the EPA provides funding, requests for changes by the EPA are often disregarded. Need to identify the incentive/disincentives to offer state partners for changes needed. For states building new systems, it is easier to request changes but for existing systems there is a cost factor to implement changes. Need to offer incentives or have to make it mandatory in order to get changes made” - PMO Director

“Many of these problems originate from inefficiencies in our data system management. The inefficiencies can be grouped into roughly four categories. First, inefficiency in making funding decisions. This can be remedied by system managers and stakeholders better communicating the value of systems to decision-makers.” - Questionnaire respondent

Data Stewardship

Data Stewardship, at the highest level, involves formally entrusting business responsibility to individuals (Coordinating Data Stewards – see below) who will steward a data subject area (across an enterprise) to ensure the data and metadata is of high quality, is properly defined, secured, understood, and effectively utilized.
Data Governance and Data Stewardship are closely related – they are the two primarily organizational and business oriented aspects of EDA/EIM. Data Governance makes enterprise level decisions – Coordinating Data Stewards oversee implementation of these decisions from a data and metadata perspective, and also provide guidance to the Data Governance board.

There are some examples of successful Data Stewardship in the EPA (e.g. for geospatial and facilities data), but it is often an ad-hoc and unofficial task which may be performed when time permits. Concern was raised that data stewards don’t receive the recognition or resources necessary to be effective. The term “Data Steward” is used extensively, but in many cases the use of the term does not coincide with industry practice and often means different things to different people within the EPA. In some cases (e.g. SIS/READ, data stewardship roles are finely tuned, with there being multiple levels of data stewards and each type of data steward having a distinct role name.

For this assessment, the role name “Coordinating Data Steward” is considered the individual who is the primary point of contact for a data subject area (e.g. facilities, geospatial, substances) and who collaborates with business leadership to ensure business requirements are met and coordinates with others across the agency (e.g. business analysts, Technical Data Stewards, system owners, etc) to ensure that the data in the data subject area “is of high quality, is properly defined, understood, secured, and is effectively utilized”.

“When talking about data publishing and metadata, it all comes back to governance and stewardship. Until the Agency identifies, hires, and respects data stewards (and given the proper pay), investments will be put forward which are wasteful because of redundant efforts and because the business wasn’t involved upfront.”
- Segment Architect

“All data ends up as a product or a service. The EPA is not doing a good job of governance and stewardship over the data, or even prioritizing the data that needs to be collected. There are many redundant data collections that have taken place which could have been avoided with proper governance/stewardship if there was a body to talk about the negotiations going on with the state and federal government to consume the secondary data.” - Segment Architect

Data Quality

Data Quality was a key concern raised during most interviews and questionnaire responses.
“The Obama administration’s commitment to transparency has ramifications, especially with respect to increasing the imperative for high levels of data quality, and is a key business driver of change.” Branch Chief (Acting)

“Data quality should be treated as a business issue – instead it is handed off to IT to handle.” Segment Architect

“Data Quality isn’t a concern – until it becomes a problem!! Data is often held tightly because don’t want to release bad data. Regions would often call the states because the national system’s data quality was so bad”– Segment Architect

“Problems with data quality are not transparent to the EPA or even to the Office of … leadership. There is a lot of effort to try to clean up data from a state, but the awareness of the effort involved isn’t raised up to leadership. These problems happen all the time.” Segment Architect

In many cases, the EPA receives data created by states, tribes, local governments, industry, and labs. However, the EPA is often blamed for poor data quality stored in its systems.

“…data quality is always an issue. People assume that the states know what to put into an EPA database. Since it is an EPA database, the EPA is held responsible where data quality issues may come from the source.” - Segment Architect

“Every segment is faced with data quality issues from the data sources. The EPA doesn’t often have control over data sources – there are often loopholes in regulations that states sometimes use.” - Segment Architect

“Regions would often call the states because the national system’s data quality was so bad.” - Segment Architect

“There are huge issues of who is responsible for the quality of the data - if the states (or local governments or regulated entities or grantees) report the data, aren’t they responsible for the quality? … We have been burned before by reporters submitting incorrect and even misleading data that the public then blames us for having in our databases.” - Segment Architect

In some cases, data quality issues are corrected by the EPA in the process of populating its data stores, in other cases the data quality issues are raised to the awareness of the states (or data creators) who may or may not correct the problems. This points to issues of governance, communication, providing incentives/disincentives, and prioritizing data quality issues to be addressed.
“Problems with data quality are not transparent to the EPA or even to the Office of … leadership. There is a lot of effort to try to clean up data from a state, but the awareness of the effort involved isn’t raised up to leadership. These problems happen all the time.” Segment Architect

“We also spend tons of money in programs like TRI to correct the data.” Segment Architect

“The regions were tasked to identify where the issue (with enforcement sensitive data) occurred and within 2.5 weeks the data was cleaned up. This was a massive cleanup effort.” Branch Chief (Acting)

As indicated above, communication regarding Data Quality issues is intermittent at best. There seems to be limited tracking of data quality issues to see how data quality is improving (or not) over time, or to identify the source of the data quality issue.

“The quicker the EPA can expose to the public the data quality issues, e.g. via a dashboard, the more the states would have to react.” – Segment Architect

“If we don’t know what the data will be used for, then how can we be sure we are collecting data at the right level of granularity, quality, and with the right metadata at the input side?” – Segment Architect

“The Environmental Working Group (a non-profit organization) published a report in the Washington Post on drinking water. They collected raw drinking water samples from almost every state, warehoused the data, and analyzed where programs were working well and where they could be improved, what the EPA could do to improve. They were a secondary user of state information – they did quite a bit of QA, and found that the utilities were very responsive to correct information. Sometimes it is just a matter of asking states/utilities to correct data, and setting up the mechanisms to facilitate this.” - PMO Director

**Data Sharing and Collection**

Data Sharing (externally and internally to the EPA) is a major concern addressed by most interviewees. The push by the Obama administration for transparency has led to concern as to the best way to share data while protecting data which should not be shared, e.g. PII (Personally Identifying Information), CBI (Confidential Business Information), radiation monitoring sites, etc.

“Data Quality isn’t a concern – until it becomes a problem!! Data is often held tightly because we don’t want to release bad data.” - Segment Architect

“The push to transparency leads to new and different data issues - puts an imperative on data quality, using the system consistently (not having creative interpretations of data definitions).” - Branch Chief (acting)
Data sharing problems are seen as a significant hindrance to the EPA accomplishing its mission; however a common belief is that by being able to share data more easily, the EPA would be more successful by having information available to stakeholders in a timelier manner.

“Consumers want to use their own toolsets and platforms – not just/even the tools the EPA provides. They just want the data and the metadata now – they are hungry for this information. If there was a terrorist attack at an intake facility the EPA wouldn’t be able to quickly report it because of the layers of business processes involved. If data was released more quickly, the reaction could be much quicker. During Hurricane Katrina, there were several problems about getting the water quality samples from facilities in real-time” - Segment Architect

The EPA provides data to the public and exchanges data with partners through a variety of means, e.g. Internet (including portals), data.gov, and the Exchange Network. Central Data Exchange (CDX) is the EPA component for transmitting and receiving data from partners through the Exchange Network, however CDX is also being used for internal data sharing as well. Some issues regarding internal data sharing were raised as well. The EPA is well along the path of a Service Oriented Architecture (SOA) using Web Services standards (HTTP, XML, SOAP, UDDI, WSDL), and this seems to be working very well to exchange data between the EPA and external parties. However, CDX seems to be a requirement for internal data exchange as well within the EPA firewall. In some cases this is hampering internal data sharing – there is programming effort required to SOA-enable an application or data integration point - more so than enabling database connectivity behind the EPA firewall. Web Services provide a powerful capability for application integration and data exchange across firewalls and for loosely coupling technology, but may not be the best solution in all instances.

“We want to make sure that all databases can talk to each other; don’t want to preclude the ability for the rest of the agency to use the information.” - Interviewee

“A headquarters pain point is dealing with chemicals and facilities using CDX. They have to go outside of the firewall and come back in using CDX web services. Most of the systems are using Oracle – yet they are forced to use CDX, which can cause issues as some don’t have excellent internet connection speed. They are considering replicating/storing the CDX data rather than interacting with CDX in real time. SRS has chemicals and substances – should be opened up for anyone in the agency to write queries against. SRS and FRS are mostly public data.” - Segment Architecture

“CDX interface has to be used, but is not user friendly and is for batch oriented, peer-to-peer XML data transfer.” - Branch Chief

In some case, data is replicated from one database to another, usually using Informatica (ETL tool). This sometimes leads to problems of data quickly becoming out of sync. There isn’t currently an Enterprise Information Integration (EII) tool in use for federated queries where multiple sources can be queried in a single query by means of a metadata layer. Instances where databases were directly connected (e.g. Oracle db instance to another Oracle db instance) were not uncovered in this assessment, and this option doesn’t appear to be utilized currently for internal data sharing. Oracle Business Intelligence is doing some federated queries for “Cleanups in my Community”.

http://semanticommunity.info/EPA/EPA_Data_Architecture/Enterprise_Data_Architecture_Strategic_Assessment_Architecture_and_Roadmap

Updated: Sat, 19 Sep 2015 03:01:20 GMT
Powered by mindtouch
The EPA in many cases does not create data but rather receives it from states, tribes, industry, etc via direct user interfaces, batch feeds, or combinations of these (e.g. RCRAinfo). Issues with data collection were a common concern.

Primary information issue: “Most data comes from secondary sources – states. Our hands are somewhat tied, re standards, quality, governance, etc – because it is from a secondary source.” Segment Architect

“…states are already combining their environmental reporting data, but then they have to separate it out again when reporting to EPA - and then EPA spends money putting it back together again. This is wasteful and inefficient.” Segment Architect

“Labs are sick of sending data repeatedly” Segment Architect

“On the input side: who is providing data to us? Do they have to provide the same data to multiple federal entities (or even multiple entities within EPA)? If so, are there any opportunities for greater efficiency?” Segment Architect

The EPA is currently unable to produce an Information Supply Chain analysis (without a lot of manual effort) which shows how data is ingested into the EPA, how data flows within the EPA, and how data is published or shared to external entities.

**Metadata Management**

The importance of metadata to aid in data sharing and data management is well recognized in the EPA and efforts to create enterprise metadata registries and repositories (e.g. SIS/READ, RCS, GDG, EIMS) have been put in place and are continuing to be developed and made available.

“The front-burner issue with data right now is how to manage the climate change data, e.g. who has it, who needs it, how it needs to be delivered, etc.” - Segment Architect

“. being able to integrate data from various sources without becoming an expert in each. The discovery problem can be remedied by inventorying and cataloging our data and service offerings and making this library available in an easy to use format” - Questionnaire Respondent

However, many issues that arose in the interviews point to problems of metadata management.

One key issue which can be addressed (in large part) by having integrated metadata is being able to handle change effectively and efficiently. One of the interview questions was “How rapidly can your organization respond to required changes to your information systems? For example, if the length of a data field had to change across the enterprise (e.g. Y2K changing 2 digit year to 4 digits), how difficult would this be?” Below are some responses to this question.
“Changing systems is hard. AFS (Air Facility System) is currently shut down in order to change the action tracking number from 3 to 5 digits. This legacy system is on a mainframe, and but there are no resources to modernize it. AFS is languishing on its current platform – the 3 digits have been used up so they have to go to a 5 digit number. The EPA is not nimble to make change. Systems are big and have many stakeholders (e.g. ICIS) – making changes is not done lightly or quickly.” – Branch Chief (Acting)

“NOTE – we need to be honest that we are not able to respond nimbly, we tend to use old business practices for development of IT systems that do not allow quick response.” -Questionnaire Respondent

“In the future, would like to be able to stick to an aggressive timeline to update entire portfolio – rather than being forced to maintain legacy systems. Recovery act showed that EPA is not nimble as a result of the legacy systems.” - Information Management Officer

“Handling changing field lengths ranges from difficult to easier. The new financial system the agency is going to has a new account code structure – every system which talks to the financial system will need to adapt to the new structure. The approach some are taking is to have a translation as an intermediary between the old and new structure/values.” - Information Management Officer

Another key issue which having integrated metadata can help with is in the area of Knowledge Retention. The EPA relies heavily on contractors and much knowledge is lost as contractors turn over, or when EPA employees transition.
“An underlying problem is that a lot of institutional knowledge is retained by the contractors.” - PMO Director

“It is very daunting when the DA doesn’t know the lines of business well as contractors might. EPA hasn’t matured to respect the cost of reliance on contractors.” – Segment Architect

“When EPA personnel (e.g. GS7’s, GS9’s, GS11’s) do get involved in detail, the employees often don’t stay in that area for long, and then contractors have to get involved again. The learning curve is substantial. The concept of knowledge management, having registries, and areas for documenting metadata to help in the exchange and managing the data on the backend has to have a direct utility to a front-line manager where they can have an analyst directly answer the questions rather than having to rely on a contractor. This is both a skills and data issue. Tools, data, and metadata need to be given to the right people to enable them to do what they need to do.” - PMO Director

“… relies upon subject matter experts for legacy system understanding. With older systems, there are only a few people with knowledge of these systems” – Information Management Officer

“Steep learning curve to gain familiarity with several systems to acquire information and/or perform data operations.” - Questionnaire respondent

“When grant money is given to the states, they hire contractors – but when the money dries up there is no one in the state who knows what the contractors have done. There is very high turnover for state technical employees.” – PMO Director

There are currently four (4) main enterprise level metadata repositories and/or registries in the EPA identified in this assessment. This does not include the many local metadata repositories or tool based metadata repositories (such as for Informatica, Business Objects, Oracle Business Intelligence, etc). Each of these four below has a particular emphasis, though there is some overlap (which is not unusual).

The following enterprise level metadata repositories were identified:

3. Quoted text is from EPA internet websites, unless otherwise noted.

### Table 1 - Enterprise Metadata Systems/Registries/Repositories

<table>
<thead>
<tr>
<th>Repository/MDR Application acronym</th>
<th>Formal Name</th>
<th>Description</th>
<th>Types of metadata stored</th>
<th>Population / Integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIS (aka READ)</td>
<td>System Inventory Services READ – Registry of EPA Applications,</td>
<td>Primarily a system portfolio repository, but has some environmental models (not data models in the DA sense). Only 9 datasets.</td>
<td>Systems/applications Environmental models. Related EA artifacts, technology used, owning/using organizations.</td>
<td>Has a front end where assigned data stewards can CRUD (create/read/...)</td>
</tr>
<tr>
<td>Models, and Data sets</td>
<td>update/delete metadata in SIS/READ, based on the users assigned data steward type. Public internet access for reading.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GDG</td>
<td>“Providing an automated connection to the central catalog removes metadata maintenance burden of the central catalog from Geospatial Data Stewards by ensuring that their GDG records are automatically refreshed along with their local records.”</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeoData Gateway</td>
<td>Geospatial metadata</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“The GDG is EPA's central geospatial metadata access point. It is an enterprise application that stores metadata about geospatial assets maintained by Geospatial Data Stewards across the Agency … The GDG is comprised of a set of Web-based interfaces and a geospatial metadata catalog”</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RCS</td>
<td>“System metadata (pulled from READ). Data set metadata (pulled from expanded GDG or other sources). Data flow metadata (pulled from ENDS). Web services metadata (populated in RCS since there is no centralized metadata source). Schema metadata (populated in RCS since there is no centralized metadata source)”. Additional metadata envisioned to be harvested (READ, GDG, ENDS, other). Direct user entry (web services metadata, schema metadata).</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reusable Component Services</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
EIMS
Environmental Information Management System

“EIMS is a repository of products and metadata. The descriptive information in metadata enables users to evaluate and use these products. EIMS stores and maintains descriptive information in a relational database and refers to the products (data, documents, etc.) stored either within EIMS or as distributed external files.”

“data sets, databases, documents, models, multimedia, projects, and spatial information”

Direct user entry

There seems to be some confusion around the terms registry and repository. A registry is a data store where data/metadata instances points to data or metadata physically stored elsewhere (e.g. in a repository). A repository is a place where data or metadata is physically stored.

RCS is moving in the direction of being the central agency metadata repository for integrating metadata from other registries/repositories. RCS uses a COTS software package called CentraSite, While marketed as a SOA governance tool, it has an extensible metamodel and has federation capabilities. Currently in RCS, metadata is being physically integrated but plans for using CentraSites federation capabilities are envisioned.

SIS/READ is an excellent tool for understanding the EPA system portfolio. As the saying goes “you can manage what you can’t measure”. SIS/READ provides descriptive information about a system and includes extensive related EA metadata about the system, (e.g. owning/using organizations, management initiatives, statutory bases, technology used, and much more). One office encourages new employees/contractors to utilize SIS/READ in order to gain an understanding of its systems.

The current emphasis of metadata repositories/registries seems to be primarily on finding and reusing existing metadata for faster application development, system understanding, and system portfolio management. Additional types of integrated MDR (in this document refers to Metadata Repositories/Registries) enabled analyses/capabilities do not seem to be currently supported such as data lineage analysis, change impact analysis, data rationalization 4 – though the need for these were expressed through the interview process.

4. See: Real World Decision Support article on Data Rationalization: http://tinyurl.com/DataRationalization
“The front-burner issue with data right now is how to manage the climate change data, e.g. who has it, who needs it, how it needs to be delivered, etc.” - Segment Architect

“I really think data flows are a very critical part of data architecture. Beginning with the end in mind - why are we collecting the data? What answers are they seeking to what questions? How will they use the data? Where will they need to use the data? If we have to collect certain data anyway, how can we put it in the most useful format, combined with the most relevant information for decision-making? If we don’t know what the data will be used for, then how can we be sure we are collecting data at the right level of granularity, quality, and with the right metadata at the input side? On the input side: who is providing data to us? Do they have to provide the same data to multiple federal entities (or even multiple entities within EPA)? If so, are there any opportunities for greater efficiency?” - Segment Architect

Data Modeling

Data Modeling currently seems to be limited primarily logical and physical data modeling of application data structures. CA ERwin is the primary data modeling tool, though Troux Metis will be used for some data modeling activities as well.

An Enterprise Data Model (EDM) (beyond the high level Data Classification Reference Model (DCM)) has not been developed which would provide an enterprise view and understanding of key agency data entities, relationships, and attributes. An EDM will be essential to helping the EPA realize its vision of data sharing and interoperability.

The DCM is what in industry is known as a “Subject Area Model” or SAM. The DCM was developed in 2002 and so it may need review and updating. It is uncertain whether the DCM was formally approved and adopted – though it is currently being referenced in SIS/READ. SIS/READ utilizes the DCM to identify the DCM levels (see below) which are affected by the system, which is an excellent EA practice.

The DCM can serve as the foundation for an EPA Enterprise Data Model (see recommendations section) pending review, updating if necessary, and formal approval and adoption. A SAM is very useful as it begins (it is the highest level data model and should be the first developed) the process of delineating or decomposing an enterprise into high level data subject areas – which is crucial for having an office/application/technology neutral view of a EPA data resources. A SAM also requires enterprise collaboration in order to achieve the semantic reconciliation needed to determining the name and definition of an enterprise data resource - and so developing a SAM and obtaining agreement and approval for it is no small feat. As the saying goes “The more often a term is used, the less it understood”. A SAM is also necessary for identifying “Coordinating Data Stewards” for the EPA.

The DCM is broken into three levels – Data Area, Data (Information) Class, and Data (Information) Sub-Class. Below is an example section of the DCM.
Table 2 - DCM Example Section

<table>
<thead>
<tr>
<th>Data Area</th>
<th>Data (Information Class)</th>
<th>Data (Information) Sub-Clas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Places</td>
<td>Natural Features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Man Made Features</td>
<td></td>
</tr>
<tr>
<td>Populations</td>
<td>Humans</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-Human</td>
<td></td>
</tr>
<tr>
<td>Environmental Stressors</td>
<td>Resistance/Chronic Stressors</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Incidents/Spills</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Habitat Depredation</td>
<td></td>
</tr>
<tr>
<td>Substances</td>
<td>Chemicals</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Biologicals</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Contaminants</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wastes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Radionics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Commercial Products</td>
<td></td>
</tr>
<tr>
<td>Facilities</td>
<td>Facilities</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maintaining Stations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other Properties or Sites</td>
<td></td>
</tr>
<tr>
<td>Sources</td>
<td>Mobile Sources</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Point Sources</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non Point Sources</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Equipment Process Sources</td>
<td></td>
</tr>
<tr>
<td>Environmental and Human Health Impacts (env)</td>
<td>Substance Exposure</td>
<td>Environmental</td>
</tr>
<tr>
<td></td>
<td>Body Burden</td>
<td>Human Health</td>
</tr>
</tbody>
</table>

Data Modeling practices are a critical component to be considered in any Data Architecture. Without consistent data modeling practices, it is very difficult to have interoperability and improved data sharing (e.g. if different offices/applications are naming the same thing in different ways with different natural keys - translation/conversion processes are required) and communication/understandability will be hindered.

There currently aren’t any EPA agency level data modeling standards. OSWER has an extensive document (OSWER Directive9028.00) on data modeling practices.

Data Models, apart from the DCM, were reviewed only at a high level due to the strategic nature of this assessment and only data models which were publicly available on the internet were reviewed. In some cases, the notation used was not as expressive as could be hoped for (e.g. FRS, LRT shows only mandatory or optional one-to-many relationships in its logical data model – but does not indicate whether a mandatory relationship is an identifying relationship (e.g. primary key of the parent becomes part of the primary key of the child). In the STORET data warehouse model, use of class words/representation terms (see ISO11179) could be added to attribute names for improved understandability.

**Data Warehousing / Business Intelligence**

The EPA has many decision support environments, e.g. Data Warehouses, Data Marts, Operational Data Stores (ODS) – but does not have a primary Enterprise Data Warehouse to provide an agency level view of data which cross-cuts offices and applications.

Many of these decision support environments are setup to meet the need of a single regulation, or even a single reporting requirement of a regulation.
“They have hundreds of regulations, but each may have separate reporting requirements, and even a single regulation may have many reporting requirements. The characteristics of one regulation might not be the same as a characteristic of another regulation. So there are hundreds of reporting requirements and hundreds of systems which are involved in producing these reports. To address these reporting needs, often the first response from program managers is “I’ve got a regulation, and I’ve got a year and half – I’m going to build my own system, and it is going to be faster and more efficient (from the program manager point of view) to stand up a single system to regulation reporting requirement - but it isn’t efficient from an EA perspective and is a mess, and puts a burden on the reporting community.” - Segment Architect

Data integration (physical or virtual) is the foundation for enabling cross-cutting DW/BI. The number of current stovepipes makes it difficult to have an integrated, holistic view of data.

“There is an ongoing need for data integration. For example, for a geographic sector (e.g. Chesapeake Bay) or an industry, there is a need to be able to measure compliance across all statutes, across all parties – taking human health risk and demographic information into account. This would be a major business driver for senior managers at the EPA – this would help them to measure if the EPA is successful and in compliance. For mid level managers, being able to know if the EPA is performing efficiently would be a key business driver.” - Branch Chief (acting)

“A great opportunity where Data Warehousing makes sense is to align financial, grants, admin, activity, and environmental outcomes data” Segment Architect

“Highly skilled (in software such as SAS) analysts generally do not use “integrated” systems.” Questionnaire Respondent

There are different Business Intelligence (BI) technologies in use (e.g. Business Objects, Oracle Business Intelligence, SAS) that seem to be used to good effect in some parts of the agency. Oracle Business Intelligence is used in some cases for querying multiple data sources. However, it appears that other areas of the agency could benefit from improved BI tools (e.g. dashboards, balanced scorecards) or improved application of existing tools/technologies.
“With executive dashboards, we have gotten better giving summary information. There is still a lot of manual data work involved in some cases. They are able to get the data they need, but when a new report is needed – you almost need to start a new project. They are able to generate reports in-house using existing BI tools.” Information Management Officer

“There isn’t currently a dynamic executive dashboard for EPA leadership. Many AA’s have a quarterly and annual dashboard, but it seems to be largely static…. OCFO (Office of the CFO) has a new high level management dashboard … CPIC also has a dashboard.” Branch Chief (acting)

“It would be extremely helpful to have useful, easy-to-design, easy to use Business Intelligence tool that provided ‘dashboarding’ capabilities…” IT Management Specialist

Structured Data Management

There is large degree of standardization on the Oracle RDBMS, though there are other database technologies used, e.g. mainframe, SQL Server.

Data silos are a common concern which results in increased costs (hardware, software, personnel), and difficulties in information sharing and interoperability. There are some efforts underway to modernize and consolidate systems (e.g. Superfund Enterprise Management System (SEMS))

“There are cases of redundant collection of data, and there are many data and metadata silos. There are many similar business processes. For example, with the NHD (National Hydrography Dataset) there are multiple collection efforts and similar processes.” PMO Director

“They have a process flow which shows that sometimes a purchase order requires data entry three different times in three different places” Information Management Officer

The EPA appears to have good practices in place regarding XML and SOA. For example, XML documents are validated against a schema (XSD). Data Exchange Templates (DET) are crafted to provide documentation for XML data elements in an easy to reference format.

Unstructured Data Management

The EPA realizes the importance of being able to find and use semi/un-structured data (e.g. data in documents, web pages, wikis, email, images, etc) effectively. The EPA has a good understanding of how semantic technologies can be leveraged to increase the find-ability and usability of data and information – regardless of how it is stored. Terminology Services utilizes a tool called Synaptica KMS for the maintenance of taxonomies, thesauri, glossaries, and keyword lists – though these appear to be in early stages of development and use.

The EPA understands that implementing semantic technologies has to be implemented in manageable iterations.
A concern raised was the numerous tools that can be used for collaboration, which causes confusion.

“There are too many competing efforts in the agency with respect to online collaboration tools. We have access to Oracle tools like the EPA Portal but also Sametime/Quickplace. And there is TeamSpace (or something like that) in Lotus Notes. None provide the fully online collaborative suite of tools that is needed to fully embrace online collaboration.” IT Management Specialist

“Currently, whenever some training occurs – a Lotus Notes database might be spun off. There is a lot that the HR system could do, but hasn’t been funded at the level needed.” Information Management Officer

The EPA has decided to standardize on the Documentum tool for ECMS (Enterprise Content Management System). A complaint was raised about the process of selecting Documentum.

“Documentum does not work well with the Portal. This is critically important since some of our applications need to be able to use both. Also, there were no web requirements considered for the content management technology decision. Documentum is for document management – not web content management, yet it is being required to be used for web content management as well”. – Segment Architect

Master/Reference Data Management

The EPA has two primary master data registries currently – FRS (Facilities Registry System) and SRS (Substance Registry System) which are part of the System of Registries (SOR). The Data Foundations OneData COTS toolset is being used to manage these MDM systems.

These master data registries provide a central integration point for the EPA enterprise for these commonly used concepts. These hubs provide for greater standardization (e.g. there are FRS interfaces with CDX where facility data from external partners can be received) as these hubs are centrally managed.

In the past at least, effective data governance and stewardship seemed to be in place for the Facilities data subject area (e.g. FRS Data Stewardship Conference, 2000).

There are many places throughout the enterprise where facility related information is used. There are many different terms used (often based on different regulations) which may refer to a facility – e.g. site, property, regulated entity. FRS realizes that enforcing standard naming across the agency will not be possible, and so allows a system to utilize the term that is most applicable.

“FRS is used for one uniform lat/long – facility id is a common identifier for a facility, generated in FRS. Their source systems feed FRS, “Cleanups in my Community” read from FRS. In FRS, there is a mapping between a source system identifier and the facility id. There is also a preferred lat/long. The facility id is not fed back into the source systems.” - Staff Interviewee
FRS and SRS seem to follow a hub approach to Master Data Management. Data about facilities and chemicals is sourced from outside the hub. Once the data becomes loaded into FRS/SRS it becomes the Authoritative Data Store (ADS) for these concepts. Data from FRS/SRS can then be pulled by external partners (e.g. request substances data from SRS via CDX) or used for other EPA systems (e.g. “Cleanups in my Community”).

There are plans to store reference data code sets in READ. Standardizing codes and code meaning will play an important role in increasing data sharing, interoperability, and integration.

There doesn’t appear to be any MDM software in use to manage master/reference data. FRS and SRS are internally developed MDM repositories/registries.

Other types of master data which are being envisioned for an enterprise MDM approach include Programs, Assets, Partners, and Regulations.

### Key EPA EDA/EIM Strengths

- An effective Enterprise Architecture (EA) program is in place with participation by a broad cross-section of the Agency. The EAWG program is supporting the Target Enterprise Data Architecture.
- The EPA seems, in general, to have solid Information Security in place (in some cases, the security was thought to be too stringent).
- FRS and SRS, master data registries in the System of Registries (SOR), aids the EPA with increased interoperability. Using mappings (e.g. between facility_id and source system facility identifiers), the EPA can have a broader “line of sight” for these key concepts across multiple systems, and also have an enterprise standard for these topics for new systems.
- The SIS(aka READ) is a useful meta data repository which provides excellent information about the EPA’s systems and is ahead of many other government and corporate enterprises in similar (if any) capacity.
- The SOR contains both metadata and master data registries/repositories and appears to be is based on architecturally sound principles.
- The EPA understands the importance of semantics in information management. The EPA has purchased Synaptica KMS to manage terminology (glossaries, taxonomies, glossaries, etc) through Terminology Services. This will help the EPA to take advantage of semantic technologies for improved interoperability, data sharing, information finding, and automation.
- The EPA is using Web Services and is continuing to advocate and implement new Web Services for data exchange between the EPA and its partners across firewalls. The EPA is also using Web Services internally for data sharing and application integration.
- The National Environmental Information Exchange Network (NEIEN) seems to be overall a robust platform for data sharing between the EPA and states/tribes/industry using XML and Web Services. Data exchanges seem to be well documented (e.g. DET’s, XSD’s) in RCS.
- The EPA has largely standardized on Oracle as the primary RDBMS platform in use at the EPA, which allows for greater standardization, better licensing opportunities, and simplified skill set requirements.
- The EPA uses industry leading COTS tools for ETL (Informatica), BI (Business Objects, Oracle Business Intelligence (OBI), SAS), Master Data Management (Data Foundations OneData, OneData MetaMap), Enterprise Architecture (Troux Metis).
Key EPA EDA/EIM Opportunities for Improvement

- Data Governance was a common concern identified in this assessment. There isn’t a single enterprise board concentrated solely on data and information (independent of office, application, or technology).
- Data Stewardship currently is performed mostly on an ad-hoc, informal basis. Data Stewards typically are not formally recognized.
- Enterprise information initiatives seem to struggle with ongoing sustenance without effective sponsorship.
- Communications could be improved to demonstrate to program offices the value that enterprise information initiatives provide. For example, maintaining SIS/READ is perceived by some as “feeding the elephant” without receiving benefit in return. However, in another case – one program office recommends that new employees/contractors utilize SIS to gain understanding of the program office.
- There currently isn’t an Enterprise Data Model (beyond the high level Data Classification Reference Model - though this is a good start). There are some EA segment CDM’s which might be useful in creating an Enterprise Data Model.
- There currently isn’t an agency level Data Modeling Standards document.
- Many data quality issues arise from external partners, yet the EPA is held responsible for the data quality in its systems. This was a common complaint and it appears that Exchange Network governance is not addressing this issue adequately.
- Data quality monitoring seems to be an ad-hoc effort. Mechanisms that provide ongoing, automated measurement and notification of data quality issues are not in place, at least not on a broad scale.
- Use of a Data Profiling tool was not expressed in this interview. A Data Profiling tool can help identify potential data quality issues effectively and efficiently - and is an ideal tool to use prior to integrating data, as well as for periodic data quality checking.
- Some have complained about being required to use Web Services / CDX, even for internal data sharing. The desire to be able to connect databases internally was expressed.
- In the draft Target Enterprise Data Architecture, four methods for data sharing were identified along with strengths and weaknesses each. These methods are ETL, EAI, Web Services, and Point to Point interfaces. However, one technology which was not included is Enterprise Information Integration (EII) which provides for federated query across a variety of data sources.
- There are many decision support environments (e.g. Data Warehouses, Data Marts, Operational Data Stores (ODS)) – but there isn’t currently a primary, agency level Enterprise Data Warehouse (EDW). A centralized, Enterprise Data Warehouse (EDW) should be considered to help reduce decision support stovepipes and improve Business Intelligence offerings.
- There is a lack of on-line Business Intelligence options available to EPA leadership (e.g. executive dashboard utilizing current information).
- There currently isn’t a formal strategy in place for providing the long term vision for RCS. There currently aren’t any plans for RCS to integrate or link to Data Warehousing meta data sources (Informatica, Business Objects, Oracle BI) or data models (ERwin, Troux).
- The current practice of setting up a new system when new regulatory reporting requirements are issued is inefficient and expensive.

Recommendations
Overview

As part of developing the Target Enterprise Architecture, the draft “EPA Target Data Architecture” document has been created in order to establish the vision and strategy for Data Architecture in the near and long term. The draft Target Data Architecture document, in general, is based on industry best practices and exhibits a solid understanding of the complexities involved in establishing a forward looking Enterprise Data Architecture, and related Enterprise Information Management (EIM) practices.

The DAMA-DMBOK (Data Management Body of Knowledge) Functional Framework is prominently referenced which identifies the key aspects of data architecture and data management. Notably, Data Governance is the core of this functional framework. EWSolutions played a significant role in developing the DAMA-DMBOK.

Figure 4 - DAMA DMBOK Functional Framework

EWSolutions has produced a functional framework for Enterprise Information Management (EIM) (below), which has correlation with the DMBOK, but with a different way of organizing EIM components. The EIM Functional Framework underscores the importance of Information Architecture (e.g. data modeling, information value chains, data flow diagrams) and Metadata Management, and also emphasizes how Data Stewardship activities need to span across the different EIM components.

The DAMA Framework is very similar to EWSolutions’ EIM framework. The EIM Functional Framework separates stewardship from governance (they are related, but separate tasks). EIM Services is an enterprise IT function which implements the EIM framework, under the direction of the Data Governance board and assistance of Data Stewards. Another key differences is that the EIM Functional Framework shows how Information Architecture, Metadata Management, and Information Security Management span the types of data and other management activities.
The recommendations will be broken into the following categories which largely correspond to the DAMA-DMBOK and EIM Functional Framework topics:

- Data Governance
- Data Stewardship
- Enterprise Data Model
- Enterprise Data Warehouse
- Metadata Management
- Data Quality Management
- Data Sharing
- Master / Reference Data Management

The recommendations are based on review of the draft “EPA Target Data Architecture” document, interview findings, questionnaire responses, and a high-level review of other documentation. Industry best practices were heavily drawn from when making recommendations.

The final part of this section will focus directly upon specific recommendations for the draft “EPA Target Data Architecture”.

**Functional Recommendations**

**Data Governance**

Data Governance has been correctly identified as a critical foundation for realizing the Target Data Architecture vision – without the effective sponsorship, oversight, leadership, and enterprise vision (from both a business and IT perspective)
that a Data Governance board can provide, it is more difficult to ensure participation of and alignment with the business when initiating and sustaining enterprise information management initiatives.

“There are many redundant data collection efforts that has taken place which could have been avoided with proper governance/stewardship, if there was a body to talk about the negotiations going on with the state, federal government to consume the secondary data” – Segment Architect

The key reasons for formulating a Data Governance board are:

- Data and information are recognized to be critical EPA assets. A Data Governance board can provide a central point of authority, responsibility, and accountability for these assets.
- Forum for communication and decision making to ensure better business and IT alignment for the effective management and utilization of EPA information assets.
  - Prioritization, sponsorship, sustenance, and oversight of EPA enterprise information initiatives.
  - Enterprise vision enables reduction/elimination of duplicative data, systems, infrastructure, and initiatives - with significant cost savings.
- More timely resolution of significant Enterprise Data Architecture and Enterprise Information Management issues.

The proposed name in the Target Data Architecture for the data governance board is the “Data Advisory Council” (DAC). The DAC should govern all types of data and metadata – whether environmental, administrative, financial, HR, etc, and should be independent of office, application, and technology.

There has been some discussion around adjusting the “Exchange Network Subcommittee” (ENS) to address data governance; the risk however is that the emphasis on the Exchange Network would lead to neglect of data and information needs found apart from this key application. Data Governance, to be effective at the enterprise level, needs to be independent of application in order to take all of the agency’s data and information needs into account.

It is recommended that the DAC have significant (primary) business representation. IT should have representation in the Data Governance Board – but should NOT have overriding or primary representation. When IT has primary representation – business involvement very often tends to drop off (this is a very common finding in industry).

It is recommended that the DAC be headed by a Chief Data Steward or Chief Information Steward who ideally would be a senior leader from outside of IT to help ensure broader business participation. The lack of business participation has been pointed to as an issue when making IT decisions. The Chief Data Architect should serve in a partnering role with the Chief Data Steward, to assist in program management activities and as a primary representative of IT and EA.
A project to undertake as soon as possible is to formulate a Data Governance Charter to determine composition, voting, guiding principles, key relationships, and communication plan. For Data Governance to be successful at the EPA, the Data Governance charter must be approved and endorsed at the highest level in the agency as possible.

Data Stewardship

With some notable exceptions, Data Stewardship at the EPA is primarily an ad-hoc, informal effort. It is recommended that the EPA institute formal Data Stewardship.

Formalized Data Stewardship will allow for improved oversight of a data subject area, e.g. facilities, chemicals, geospatial data, programs, partners, etc. There has been effective stewardship of facilities and geospatial data (at least) – this should be recognized, and best practices from the stewardship of these subject areas be compiled and leveraged.

Key benefits for formalized Data Stewardship include:

- Provides a central point of contact (Coordinating Data Steward) for a data subject area – improving collaboration, business/IT alignment
- Regulation stewards provide a central point of contact for a specific regulation to provide understanding of how a regulation will impact the information environment and conversely, how changes to data assets will affect regulation compliance
- Identifies opportunities for reuse (data, metadata, systems, infrastructure, data modeling)
- Improved accountability for the data subject area
- Improved data quality through ongoing monitoring of data quality for a subject area
- Data assets are better understood from an agency and business perspective (e.g. improved definitions, data model review/approval, creation of stewardship metadata)
- Improved standardization in naming and definition of data resources, for PII and CBI compliance monitored at the enterprise level for the subject area
Data Stewardship should be formalized based primarily on data subject areas, and then secondarily based on regulation, and EA segment, as EPA data frequently spans multiple regulations, program offices, EA segments. This will enable a more holistic view of data and ensure that data for a subject area is properly understood, secured, effectively utilized, and data quality is measured and improved. Having this enterprise view to a data subject area means that opportunities for reuse are improved and expense due to duplicative systems, data stores, data modeling, and infrastructure can be minimized, where possible.

“Coordinating Data Stewards” should be identified for each data subject area, and empowered with responsibility, authority, and accountability to steward the data subject area.

An excellent recommendation put forth by an interviewee was to have “Regulation Stewards”, as regulations often will span multiple data subject areas. An ideal Regulation Steward might be the foremost expert on the regulation in the EPA.

“Coordinating Data Stewards” will be assisted by technical data stewards and subject matter experts – they need to have the appropriate level of visibility in the EPA and would ideally come from outside of IT – ideally with at least a basic understand of data management concepts.

“Coordinating Data Stewards” will need to be able to interact effectively with EPA leadership and system owners and will need to be able to garner agency-wide consensus, and so must have the requisite interpersonal and organizational skills, with corresponding compensation. For “Coordinating Data Stewards” (at least), the task of data stewardship should be an activity for which they are held accountable for in their performance reviews. If not, there is risk that data stewardship will be relegated to an activity to be performed when time permits.

A Data Stewardship Sub-Committee should be formed which would be accountable to and would advise the DAC. In the Target Data Architecture the “Enterprise Data Architecture Working Group” (EDAWG) has been proposed as a subcommittee of the DAC. It is recommended that a Data Stewardship Subcommittee be separate from the EDAWG as the EDAWG’s emphasis is on Data Architecture. A Data Stewardship Sub-Committee should be an ongoing endeavor as changes to regulations and other impacts will require a sustained Data Stewardship organization.

Key benefits of a Data Stewardship Committee include:

- Determining data stewardship practices and standards
- Determining tools and templates for effective stewardship
• Review enterprise information initiatives for alignment, reuse opportunities
• Provide input on and review / approve Enterprise Data Model
• Forum to resolve issues not appropriate to the DAC

Some data subject areas may be so complex as to require a subcommittee just for the data subject area. Such a subcommittee would be led by the “Coordinating Data Steward” and may be also comprised of Technical Data Stewards, System Owners, Subject Matter Experts (SME), and other interested parties.

The Data Classification Reference Model (DCM) identifies key data subject areas in the EPA. The DCM should be reviewed, updated, diagrammed, and approved by the DAC in order to delineate the data subject areas and to help identify “Coordinating Data Stewards” for the subject area. Identifying or hiring “Coordinating Data Stewards” will be one of the earliest and most critical activities of the DAC.

“When talking about data publishing and metadata, it all comes back to governance and stewardship. Until the Agency identifies, hires, and respects data stewards (and given the proper pay), investments will be put forward which are wasteful because of redundant efforts and because the business wasn’t involved upfront.” – Segment Architect.

Enterprise Data Model

In order for the EPA to share, integrate, analyze, and use data more effectively, and enable greater interoperability across systems via the data layer, it is imperative that enterprise data be effectively defined and understood. It is strongly recommended that an Enterprise Data Model (EDM) be developed to help identify, name, and define key enterprise business concepts and data objects in a manner independent of office, application, system, or technology.

5. An enterprise name is required to improve standardization, but this does not mean that synonyms or homonyms must be removed throughout the enterprise (as this is impractical). However, noting these synonyms and homonyms in EDM metadata and in Terminology Services will be very important in order to map data across systems.

Complex initiatives, e.g. building a house, are always preceded by the development of the architecture – diagrammatically, with many overlaying models. To help facilitate the future state information environment the EDM will demonstrate what the key business concepts/data objects (not systems, databases, etc), are and how they relate.

“Unless you are aware of the larger concerns, decisions to optimize will focus on immediate concerns…”

Segment Architect

Key benefits of an Enterprise Data Model include:
• Improved alignment and communication across the enterprise (business and IT).
• Framework for development of future Information Systems, especially enterprise applications such as an EDW and additional MDM hubs.
• Enables improved data sharing due to semantic reconciliation of data resource naming (e.g. synonyms, homonyms identified, linked to enterprise standard name)

• Enables effective data stewardship

• More efficient development of application data models, DET’s, XSD’s as enterprise definition, semantic reconciliation can be reused

An EDM is actually a series of models (typically three levels) with the Subject Area Model (SAM), Enterprise Conceptual Data Model (ECDM), and Enterprise Logical Data Model (ELDM) as the most common components.

The EPA has a head start on developing an EPA Enterprise Data Model given the work done with the Data Classification Reference Model (DCM), which is equivalent to a Subject Area Model (SAM). The DCM should be reviewed, updated, and diagrammed (for faster comprehension). In order to identify Coordinating Data Stewards, the DCM should be approved and promoted by the DAC. Also, Segment Architects have been tasked with developing a Conceptual Data Model (CDM) of their segments. Both the DCM and segment CDM’s can be leveraged when developing the EDM. The EDM, in order to be effective, must be approved and promoted by the DAC, with significant prior review and approval by the Data Stewardship Committee.

**Figure 8 - EDM levels**

**Figure 9 - Example Subject Area Model (for Healthcare)**

http://semanticommunity.info/EPA/EPA_Data_Architecture/Enterprise_Data_Architecture_Strategic_Assessment_Architecture_and_Roadmap

Updated: Sat, 19 Sep 2015 03:01:20 GMT

Powered by mindtouch™
Enterprise Data Models are often developed use Information Engineering (IE – crows foot), UML, or Object Role Model (ORM) data modeling notations. Troux Metis is used for the Enterprise Architecture tool. The EDM can be developed, at least in part, using this tool. Other EA artifacts can be linked to EDM artifacts to provide a broader “line of sight”.

It is important that the EDM be diagrammatical in nature as a primary goal of an EDM is to foster business and IT alignment and improve communication and knowledge retention. In addition to the diagrams, documentation needs to accompany the EDM. One of the most difficult aspects of developing an EDM is semantic reconciliation. Terms identified in the EDM process which have not been previously defined (or not adequately) should be registered in the Terminology Services glossary. In the course of developing the EDM, many synonyms and homonyms will be uncovered – these should be captured and stored as metadata. This will be invaluable to aid Data Rationalization where EDM artifacts can be linked to implementation artifacts (e.g. an ECDM entity mapped to an ELDM entity which is mapped to an application logical model entity which is mapped to a physical database table or XML document). Without such Data Rationalization capabilities (which can be facilitated by the Managed Metadata Environment (MME) – see below) Data Stewardship will be made more difficult and it will be more difficult to effect change efficiently.

The EDM is also different than an enterprise application data model (e.g. Data Warehouse, MDM Hub). The EDM would be used to inform and serve as a reference for application data models (whether enterprise or not) to ensure better standardization for improved data sharing and interoperability.
There is concern that developing an EDM would be expensive. Successful EDM efforts are always performed in iterations, focusing on the subject area(s) which can support near term needs. For example, if a new MDM registry/repository is envisioned, data modeling is a necessary prerequisite. To support the MDM data modeling effort, the EDM can be reviewed or expanded to ensure the MDM topic is well understood before application level logical and physical modeling commence. In this manner, EDM development can directly support enterprise initiatives. Successful EDM efforts are almost never developed in a “big bang” approach – instead an iterative process is the most successful.

The EDM should be maintained over time – failure to do so will result in loss of value of the EDM. This is especially important for the EPA given the changing regulatory landscape.

The EPA could also benefit by establishing enterprise-level data modeling standards and guidelines to improve enterprise alignment, leverage best practices, and increased data standardization.

CA ERwin seems to be the primary logical and physical data modeling tool used – developing standard ERwin templates, domains, naming, and abbreviations for use across the enterprise will help improve semantics, standardization, modeling efficiency, and ultimately, more interoperable systems based on the data layer.

Data entities/attributes which will contain PII or CBI, or which has specific applicability to a regulation should be appropriately classified as such in data model metadata (e.g. ERwin UDP).

Data Models should be imported into the MME to help facilitate measurement of compliance to privacy/confidentiality and regulatory requirements so that information transparency per President Obama’s mandate can be achieved appropriately.

“Within a given area of business, people who have a need to know, know what data is CBI. It is uncertain as to how this should be communicated across the agency.” Segment Architect

Enterprise Data Warehouse

It is recommended that the EPA investigate business cases for the creation of an Enterprise Data Warehouse (EDW). An Enterprise Data Warehouse would be the agency level physical and virtual environment, where, for key types of data, data is integrated, cleansed, standardized, and enriched and is linked with other pieces of data to provide for efficient and effective strategic, tactical, and possibly operational decision making. Enterprise Data Warehouses provide significant value to many very large organizations and have scaled in some cases to near petabyte sized databases.

In proposing an Enterprise Data Warehouse, it is not envisioned that the other decision support environments would no longer be supported – though for financial reasons it may make sense for some decision support environments to be folded into the larger environment. The primary goal in recommending investigation of an Enterprise Data Warehouse is to support the EPA in making effective and efficient decisions – ideally to make decisions which have not been able to be supported in the current environment, or without significant manual intervention.

There was some interest in a “Business Warehouse” in the past.
"In the past a Business Warehouse was proposed (for 10 years), but was too early – there wasn’t the money or interest. John Sullivan was a proponent for the Business Warehouse. 5-7 year later – there are now some very obvious needs. A virtual warehouse is a possible solution – need to know what technologies to bring to what bear at the right time. There are some cases where regulatory constraints may prevent integrating some data together in a physical data warehouse. A virtual warehouse was expressed as an alternative. A great opportunity where Data Warehousing makes sense is to align financial, grants, admin, activity, and environmental outcomes data “ – Segment Architect

Integrating data, as evidenced by the above quote, has been identified as a concern due to regulatory or other constraints. There are, however techniques which can be used to support integration which can still be considered in compliance with regulations (of course, depending on the regulation). For example, deidentification is a method used when historical analysis is required but PII must be respected. Row level security is an option commonly employed to fine tune authorization so that within the same table, multiple types of users can have access to data records as appropriate to their authorization.

Successful Enterprise Data Warehouses are always based on supporting key business drivers. The primary reason for Data Warehousing failures is lack of clear business drivers. For this reason, business drivers need to be identified and documented before the decision to implement an Enterprise Data Warehouse can be made. Enterprise Data Warehouses, when supported by clear business drivers, developed in an iterative manner with experienced project managers and architects and with the requisite executive sponsorship have a well established record of success in industry. Enterprise Data Warehousing is in its maturity, and proven, established methodologies (e.g. EWSolutions I3sm ) can be utilized to help ensure success.

Some possible business drivers to consider, which arose from the interviews include:

• “A great opportunity where Data Warehousing makes sense is to align financial, grants, admin, activity, and environmental outcomes data”. - Segment Architect

• “For example, for a geographic sector (e.g. Chesapeake Bay) or an industry, there is a need to be able to measure compliance across all statutes, across all parties – taking human health risk and demographic information into account. This would be a major business driver for senior managers at the EPA – this would help them to measure if the EPA is successful and in compliance. For mid level managers, being able to know if the EPA is performing efficiently would be a key business driver” - Branch Chief (acting)

• “In the future, there is a good possibility that some systems will be hosted elsewhere, e.g. grants, payroll, time and attendance, etc. The need for Data Warehousing for integrated reporting to pull this data back will be important. Admin and financial data will need to be integrated” IMO

• “they have hundreds of regulations, but each may have separate reporting requirements, and even a single regulation may have many reporting requirements. The characteristics of one regulation might not be the same as a characteristic of another regulation. **So there are hundreds of reporting requirements and hundreds of systems which are involved in producing these reports.** To address these reporting needs, often the first response from program managers is "I've got a regulation, and I've got a year and half – I'm going to build my own system, and it is going to be faster and more efficient (from the program manager point of view) to stand up a single system to regulation reporting requirement - but it isn't efficient from an EA perspective and is a mess, and puts a burden on the reporting community. " – Segment Architect

• (emphasis added)

An Enterprise Data Warehouse could significantly aid the EPA in its mission by:

• Simplifying publishing and analytics

• Providing new analytical capabilities

http://semanticommunity.info/EPA/EPA_Data_Architecture/Enterprise_Data_Architecture_Strategic_Assessment_Architecture_a
Updated: Sat, 19 Sep 2015 03:01:20 GMT
Powered by Qmindtouch™
• Increasing and simplifying data sharing
• Gaining economies of scale (minimize redundant decision support environment and DW/BI efforts - wherever possible and feasible)

An EPA Enterprise Data Warehouse can simplify publishing and analytics by providing integrated and standardized data to stakeholders through an architecture which supports robust and high performing Business Intelligence (BI) capabilities. In some cases currently, decision makers need to navigate multiple systems in order to obtain the data needed to make decisions.

Data integration is an integral component of an Enterprise Data Warehouse. This integration can be accomplished physically in a centralized database, or virtually using federated query capabilities – though any Enterprise Data Warehouse will almost certainly need to have a significant physical data integration and storage component. Careful thought needs to be employed when determining which method is appropriate.

Virtual data integration can be accomplished using EII software (federated queries using a metadata layer), Web Services, BI tools, and by other means – where this makes sense. Performance (primarily query performance) is a key consideration to take into account. Some instances where virtual data integration will be less desirable include:

• Complex integration and transformation required – will affect query performance negatively and is inefficient as this integration and transformation will have to be performed many, many times for the same data whereas with physical data integration these actions are performed once.
• Performance impact on source system environments – BI queries often require thousands, millions, or billions of records – this can impact performance of source systems which may be tuned for a different requirement (e.g. transaction processing).
• Large numbers of source systems to integrate in a single query – query performance can be hindered. If one source system is unavailable – the query will fail.
• Large volumes of data needed in a single query (can affect network performance)

Historical data needed in order to support historical trend analysis may be purged from source systems due to differing data retention requirements. EDW’s typically store 2 – 5 years of data online.

It was expressed that there currently isn’t an online, dynamic executive dashboard for most EPA executives. Currently, mostly manually produced static reports or dashboards are produced. It is recommended that enabling online, dynamic executive dashboards can help executives make more effective and efficient decisions, with a graphical and intuitive interface. A Balanced Scorecard may be one approach to dashboarding – Balanced Scorecards use leading indicators vs. lagging indicators (e.g. accounting based) and enable decision makers to be more proactive and less reactive to changing business conditions. Balanced Scorecards are typically separated into 4 key sections: Customer, Internal Business Processes, Education and Learning (how is the organization learning and improving), and Financial.

Data Sharing can also be improved using an EDW. For example, by establishing one interface between the source system and the EDW or by storing the data in the EDW duplicative data sharing efforts can be avoided and data can be shared more rapidly.

Below is an example of what an EDW might look like, though few EDW’s will incorporate all of the components shown below.
Metadata Management

It is recommended that a formal MME (Managed Metadata Environment) long-term strategy be created in order to confirm RCS’s infrastructure, federation capabilities, and capabilities for extending the vision of metadata management beyond finding and reusing metadata (which is excellent in itself) to supporting additional types of MME-enabled applications, and to support a broad array of metadata types. A for example, to support the current Data Warehousing environment and a future Enterprise Data Warehouse RCS would need to be able to harvest metadata from DW/BI tools (e.g. Informatica, Business Objects, Oracle BI, SAS), data models, and other sources of metadata: See Figure 15 below.

A Managed Meta Data Environment (MME) ‘represents the architectural components, people and processes that are required to properly and systematically gather, retain and disseminate meta data throughout the enterprise”. David Marco “Universal Meta Data Models”

The technology RCS is using for its metadata repository is CentraSite. CentraSite, though marketed as a SOA governance tool, seems to be able to accommodate many types of metadata and has an extensible metamodel. The Data Standards Branch (DSB) has, in fact, already extended the metamodel to accommodate additional types of EPA metadata.

Key benefits to the EPA in expanding RCS as the enterprise MME:

- More effective and rapid application development and maintenance
- Improved knowledge retention. Faster system and data understanding - new employees/contractors can be brought up to speed much more rapidly
- Support system consolidation to reduce information silos, duplicative efforts
- Support new types of analysis
- Provides opportunities for increased automation.
- More robust enterprise applications
- Improved business agility – more rapid and less disruptive data/system changes
Some examples of MME-enabled applications which the EPA might investigate include:

- Information Supply Chain analysis (horizontal data lineage)
- Change Management Analysis (impact analysis)
- Data Governance/Stewardship technical backbone
- Data Rationalization (for increasing semantics – vertical data lineage)
- Automation (e.g. data standards compliance checking, data quality monitoring)

It is important for a MME to be able to support a wide variety of metadata types, and be able to provide visualization capabilities to aid improved analysis.

Without relationships between metadata instances, the value of the MME is decreased and this can lead to islands of metadata – even if the metadata is in the same repository! A robust MME environment should support direct user access to make connections between metadata instances as well as create/update/delete metadata. For example, user metadata entry screens might be developed for capturing data definitions created by Data Stewards.
A critical requirement is the need to socialize and provide training for RCS to help ensure broadest usage possible. Identify, document, and promote use cases where RCS can improve information sharing, interoperability, knowledge retention, and speed application development efforts.

To sustain any type of enterprise application, it is important to measure and monitor the usage and efficacy of the application. For an MME, this need is even greater as the business might not realize the value that the MME brings to the enterprise. For example, at a large credit card company, they were able to calculate cost savings dynamically based on activities the MME was being used for.

Determine and capture metrics for measuring how the MME improves Knowledge Retention. For example, in one office new employees are directed to explore SOR in order to understand the office and its systems.

Another recommendation is to determine and approve a standard minimum set of metadata elements/tags to be attached to both structured and unstructured data. Identify methods for attaching these metadata elements to data, determine use cases, and methods for harvesting these metadata elements into the MME. This will enable a broader line of site across the EPA for these elements. For example, assume one wanted to find all the documents created by “Lisa P. Jackson” rather than all documents which contain the name “Lisa P. Jackson” somewhere in the document. Using the Dublin Core element “Creator” might enable all the EPA to find all documents created by “Lisa Jackson” – rather than comb through all the results found by a search engine. Metadata as describe here come into play primarily with documents or other semi/unstructured data. However, these tags may also be associated with structured data in a relational database – though the number of tags implemented would probably be smaller given the amount of potential space these could consume. Of course, retrofitting some of these tags into an existing database would be difficult, but for new data stores this can enable the EPA to not only find documents created by Lisa Jackson, but structured data records created by her (of course, assuming proper authorization granted…).

**Data Quality Management**

A key concern regarding Data Quality is the quality of data received from states, tribes, and industry partners. Coordinating Data Stewards (with assistance from Technical Data Stewards) should measure and assess the quality of data received from external parties for the subject area, and make recommendations for correction.
A mechanism to collect and share (e.g. dashboard) Data Quality metrics should be established. External and internal data creators should be made aware of the quality of data that is being provided to the EPA, to incent improved data quality.

The MME should be used to store data quality metrics – and provide the lineage to identify the Information Supply Chain so as to help identify the source of Data Quality issues.

Benefits for capturing, measuring, and publishing data quality statistics include:

- External providers are incented to improve the quality of submitted data
- Data Quality can be continually and proactively monitored using automated tools and dashboards to reduce instances of reactive and costly responses
- Improved confidence in reporting to stakeholders
- Reduced costs due to improved data quality correction by partners
- Minimize “blaming” of the EPA due to data quality issues arising from partners

Federal Data Quality Guidelines has been created – it is recommended to review and adopt these guidelines where this makes sense.

**Data Sharing**

The current emphasis on utilizing Web Services is feasible for data sharing, given its ability to work across firewalls and be loosely coupled, technology-wise. However, some opportunities for increasing the amount of data sharing, and improving the speed and efficiency of this sharing might be realized by some alternative methods. For example, investigate connecting databases directly, when possible, for systems behind a common firewall. Oracle is the most commonly used DBMS, and has robust security mechanisms. Enterprise Information Integration (EII) technology (e.g. Composite, MetaMatrix) is another relatively easy way to enable data sharing through federated queries. EII software can generally access Web Services as a data source, and so a federated query could access resources within or across a firewall. EII software uses a metadata layer to hide connectivity complexities from the user. A single user query might be interrogating any number of data sources simultaneously – for example, a Web Service, an Oracle database, a mainframe database, an XML file might all be interrogated in the same query.

Identify optimal means of data sharing internally given the application requirements. Web Services might not be the best solution in all cases.

In some cases, providing a variety of means for sharing the same data may make sense to help enable the broadest data sharing possible and to efficiently meet varying needs of applications.

Consider leveraging an Enterprise Data Warehouse as a means of data sharing. This can minimize the amount of point to point data sharing efforts and reduce development efforts. Analysis work related to finding where to go to get data can be minimized using and EDW.

New systems should be developed with the ability to share data as an essential requirement (rather than an afterthought) given the emphasis on transparency. This makes good sense given the push for transparency and since the same data is often required for multiple applications. The greenhouse gas system is being designed with data sharing in mind.
Recommend investigating modification to the SLCD to emphasize data sharing as a requirement in new system initiatives. In the past, contractors weren’t incented to build data sharing capabilities into applications.

**Master/Reference Data Management**

Continue to leverage FRS and SRS, and investigate opportunities for additional master data registries/repositories. RCS is envisioning loading reference data code sets into its repository and making these available. This will increase standardization and can improve interoperability dramatically, as well as aid publishing and analytics capabilities. For example, if every system is using the same code values for US states and territories, then by default – it will be easier to share and interoperate. If however, some systems used a 3 digit state code and others used the 2 digit state code, then it is more difficult to share and interoperate as a conversion or translation process has to occur. (For illustration purposes only – hopefully all systems are using the FIPS standard for state code).

New applications, where possible, should leverage FRS and SRS as the Authoritative Data Source for these master data topics, and if possible align directly with these. For example, using the “facility_id” generated in FRS for identifying a facility as the key (primary or alternate) in new systems will increase standardization and reduce need for maintaining mappings between “facility_id” and the local identifier for the facility.

Investigate possibility of enabling direct database connectivity to enable more direct and rapid access to these Master/Reference data sources – rather than forcing use of only a Web Service for these.

Continue identifying Authoritative Data Sources (ADS) – not just for Master/Reference data. Identifying ADS’s will minimize the amount of analysis required to find and obtain the authoritative data. This will be especially helpful for an EDW effort, where a major task is to identify authoritative sources.

**Target Enterprise Data Architecture Document Recommendations**

In general, the draft EPA Target Data Architecture document provides a solid vision and strategy for the future state EPA Data Architecture and Enterprise Information Management (EIM) practices and should be considered for approval pending review of the above recommendations and consideration of the following.

Key recommendations for this document:

1. Add a guidelines section for creating an Enterprise Data Warehouse. While Data Warehousing is mentioned as a future consideration, it appears that an EDW may provide significant, near and long-term benefit to the EPA.

2. Consider refining the Data Governance section based on the recommendations above. In the “Layers in a Logical Architecture” image, add a Data Governance and Data Stewardship layer.

3. Add a guidelines section for development of an Enterprise Data Model and for development of enterprise data modeling standards (standards for data modeling which apply across the EPA).

4. Emphasize need for metadata integration (physical and/or virtual) via an enterprise Managed Metadata Environment (MME). Metadata integration will enable powerful metadata analysis capabilities. Current emphasis is primarily on finding and reusing metadata (which is excellent in and of itself).
5. Recommend refining the “Tools Used in the Metadata Process at EPA” image to include an Enterprise Metadata Repository layer, metadata integration layer, and metadata application layer.

6. Emphasize need to identify and capture data quality metrics as part of the “Data Quality Assessment Process”. Automate capture of these as much as possible, and store for historical data quality trend monitoring.

7. To align with the EPA and FEA DRM, recommend reviewing uses of the term Data Asset and Data Resource in the document. Ensure consistent use of these terms in alignment with the FEA DRM definition. Per the document, examples of a Data Asset include databases, web sites, document repository, etc. In the EPA DRM, it shows that a Data Steward manages a Data Asset – however, best practice is for a Data Steward (especially a Coordinating Data Steward) to steward (not manage) a Data Resource. Typically, System Owners manage a Data Asset.

Figure 16 - EPA DRM (partial)

8. Crafting distinct terms for different types of Data Stewards and Data Management roles (e.g. Coordinating Data Steward, Technical Data Steward, Executive Data Steward, Chief Data Steward, System Owner, Subject Matter Expert, etc) would be helpful for communication and differentiating roles.

Key Next Steps

Initial priorities

Key initial priorities for instituting the Target EDA and implementing EIM initiative are:

1. Refine, obtain approval for, and promote the Target Enterprise Data Architecture

2. Develop DAC (data governance) Charter, identify Executive Sponsor and Chief Business Data Steward, obtain approvals (QIC)
3. Develop an Enterprise Data Modeling (EDM) strategy

4. Review/refine/approve the existing Data Classification Reference Model (DCM)

5. Obtain approvals from DAC & QIC for the EDM strategy and DCM

6. Coordinating Data Stewards identified by DAC based on DCM subject areas

7. Develop an Enterprise Data Warehouse strategy

8. Form Data Stewardship Committee; identify data stewardship policies and tools

9. Develop a near and long term MME strategy

The draft Target Enterprise Data Architecture document should be refined based on the above recommendations and be submitted to the highest level approver or approving body possible in order to garner the necessary support to continue the EDA program and institute EIM initiatives.

- Post this approval, the following EDA/EIM initiatives can be started and executed in parallel.
- Develop DAC Data Governance Charter (priority 2 above), and obtain approval and support for it.
- Develop an Enterprise Data Model strategy (priority 3 above) and review/refine/approve the DCM (priority 4 above)
- Initiate an Internal Data Sharing assessment which will investigate the feasibility of using alternative methods to share data besides Web Services (e.g. direct database connectivity, EII, EDW).

**Near-term plan for EDA/EIM initiatives**

**Figure 17 - Near-term EDA/EIM Plan/Priorities**

**Key Dependences**

Developing an Enterprise Data Model strategy and refining the DCM should be completed early in the process as Data Stewardship and the EDM are dependent on these.

Data Governance, Data Stewardship are dependencies for nearly all EDA/EIM initiatives.
Having the MME in place prior to 1st EDW iteration is a strong plus, especially if metadata can be harvested from DW/BI tools (Informatica, Business Objects, Oracle, etc).

The MME should serve as the technical backbone for Data Stewardship, and so enabling Data Stewardship functionality in the MME should be an early implantation, second only to supporting an EDW.

The EDW and MDM projects are dependent upon the EDM.

The EDW partially dependent upon Data Quality cleanup – sourcing from corrected and cleansed data sources will improve the Data Quality of the EDW.

The Data Quality Assessment should not be completed before Coordinating Data Stewards assigned in order to assist and review.

---

**Appendix A: Glossary of Terms**

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>“Facts represented as text, numbers, graphics, images, sound or video. Data is the raw material used to create information.” <a href="http://semanticommunity.info/EPA/EPA_Data_Architecture/Enterprise_Data_Architecture_Strategic_Assessment_Architecture_a">DAMA Dictionary of Data Management</a></td>
<td></td>
</tr>
<tr>
<td>Information</td>
<td>Data used in context with other pieces of data to form meaningful information. “Meta data is all physical data and knowledge - containing information about the business and technical processes, and data, used by a corporation” <a href="http://semanticommunity.info/EPA/EPA_Data_Architecture/Enterprise_Data_Architecture_Strategic_Assessment_Architecture_a">David Marco</a></td>
<td>Meta data exists as information stored / contained in some medium, as well as the knowledge contained in people’s heads about data and processes.</td>
</tr>
<tr>
<td>Metadata</td>
<td>Information about data and processes</td>
<td></td>
</tr>
<tr>
<td>Metadata Management</td>
<td>The integration of meta data resources, to enable applications such as information supply chain, system portfolio management, change management impact analysis, etc.</td>
<td>A piece of data may be contained in many data stores (e.g. replication) and may be referenced by many applications, reports/cubes, interfaces, screen forms, etc. By integrating the meta data, the lineage of a piece of data can be traced - so that the data element can be better understood, secured, shared, managed, and data quality can be better assessed. When change occurs - impacted data stores, applications, interfaces, etc can be identified thus saving significant person-hours spent in manually identifying affected systems and components.</td>
</tr>
<tr>
<td>Managed Metadata Environment (MME)</td>
<td>“The managed meta data environment represents the architectural components, people and processes that are required to properly and systematically gather, retain and disseminate meta data throughout the enterprise”. David Marco “Universal Meta Data Models”</td>
<td></td>
</tr>
<tr>
<td>Data Quality</td>
<td>“The degree to which data is accurate, complete, timely, consistent with all requirements and business rules, and relevant for a given use.” DAMA Dictionary of Data Management</td>
<td></td>
</tr>
<tr>
<td>Data Governance and Stewardship</td>
<td>Data Governance entails having enterprise level (business and IT) oversight, and making strategic decisions regarding information assets. Data Stewardship entails the oversight of a specific data subject area (e.g. Facility), with regard to definition, data quality, security, and data usage. Government agencies and corporations have boards for assets such as money, people, facilities – but in the Information Age, many still do not have cross-cutting boards for the effective governance of a key asset: information.</td>
<td></td>
</tr>
<tr>
<td>Enterprise Data Modeling</td>
<td>By developing models which depict and define key data assets in a manner which is independent of business unit, application, and technology – significant interoperability, information sharing, and enterprise integration can be achieved, in a much more effective manner. Having a model of the key data entities in the enterprise requires enterprise collaboration to resolve semantic differences, but can result in: • Improved business and IT alignment as a result of improved communications and a common reference and • Increased standardization, which makes interoperability, integration, and information sharing much easier</td>
<td></td>
</tr>
<tr>
<td>Master/Reference Data Management</td>
<td>The management of common, shared data entities (in most cases) to ensure that data can be reused, have a consistent definition, consistent range of acceptable values – may involve the use of centralized repositories or registries to aid consolidation and standardization, and reduce costs due to redundant data stores, inconsistent values.</td>
<td></td>
</tr>
<tr>
<td>Data Warehousing/</td>
<td>Data Warehousing is the integration of enterprise data into a historical</td>
<td></td>
</tr>
<tr>
<td><strong>Business Intelligence</strong></td>
<td>(and sometimes current-valued) data store for the purposes of reporting and strategic, tactical and sometimes operational decision support. Business Intelligence is the use of technologies and techniques for leveraging data for effective, information-based decision making.</td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>Structured Data</strong></td>
<td>Data with a pre-defined, relatively stable structure contained apart from the actual data.</td>
<td>Example: records in a database table.</td>
</tr>
<tr>
<td><strong>Unstructured Data</strong></td>
<td>Data with no predetermined structure defined in advance.</td>
<td>Example: images, video, document.</td>
</tr>
<tr>
<td><strong>Semi-Structured Data</strong></td>
<td>Data which has both structured and unstructured data elements, or has a structure which can be easily changed by a user.</td>
<td>Example: Spreadsheet, email, image file which has meta data elements (e.g. title, subject, owner, etc).</td>
</tr>
<tr>
<td><strong>Semantics</strong></td>
<td>Study of meaning; much meta data exists (or should exist) to provide context and meaning to data and processes.</td>
<td>Examples of semantic meta data: data definitions, data examples, list of valid values, conceptual and logical data models showing data context, etc.</td>
</tr>
</tbody>
</table>