Intelligent Categorization: Air Force Information Asset Management

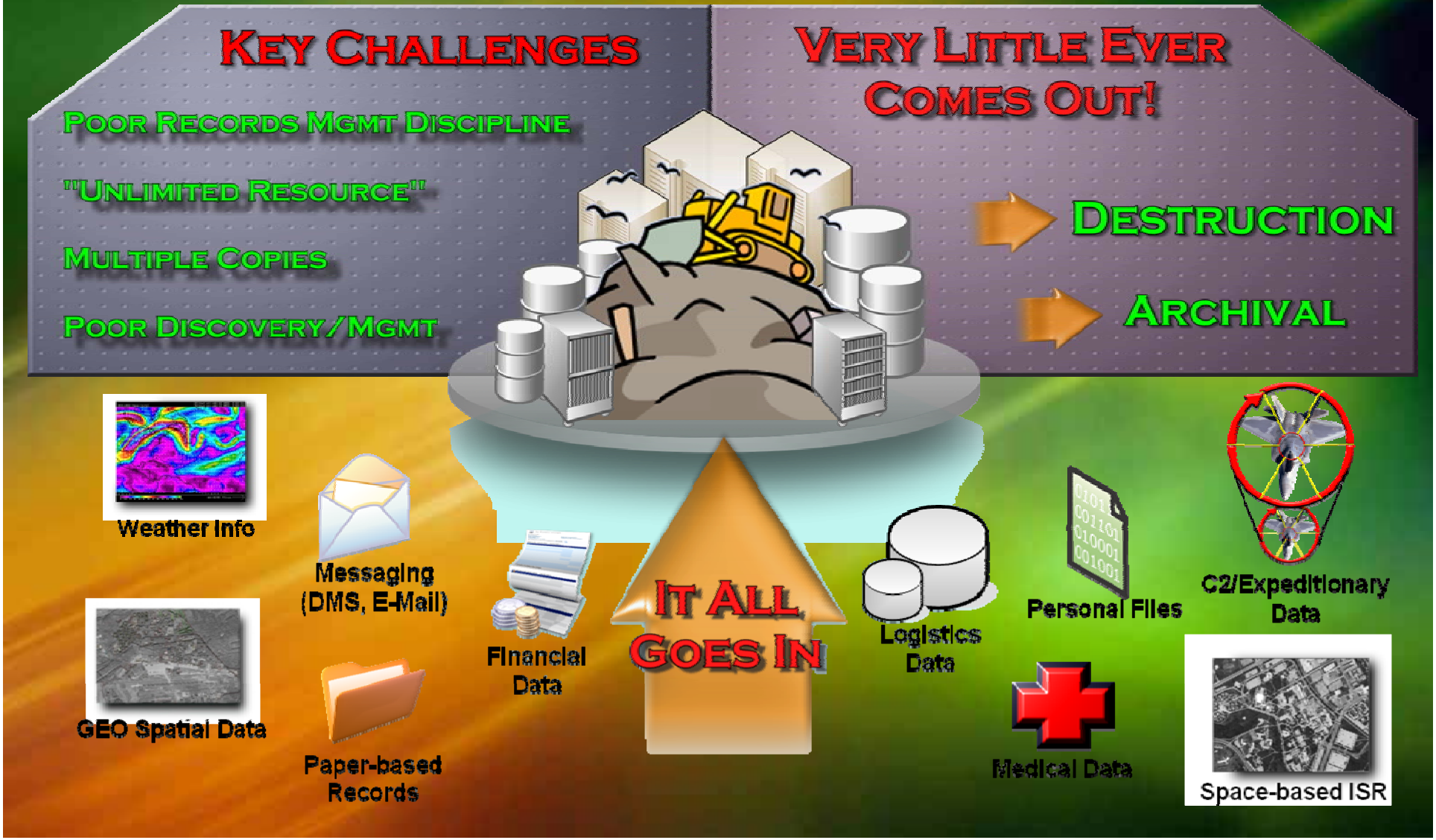
Michael Corrigan
Sumaria Systems, Inc.
Michael.Corrigan@pentagon.af.mil

J. Timothy Sprehe
Sprehe Information Management Associates, Inc.
jtsprehe@jtsprehe.co

Agenda

1. What is the problem?
2. What is the Concept for a Solution?
3. How do we implement the Solution?
4. Is this solution applicable to a broader community?

What is the Problem?
We are creating Information Landfills



We all know this problem

- Several years ago, AF realized it had over 5 petabytes of data in storage technology.
- A study showed:
  - Less than 30% of data accessed in last 12 months;
  - Less than 10% accessed in last 6 months.
  - No one could accurately identify the data or state how long it had been in storage.
  - Cost of storage: About $300M per yr and growing.

The Myth of Cheap Data Storage Technology

- People say: “Storage is cheap. Put the data in storage and move on to the next project.”
- Not when you are spending $300million+ peryear
  - Data storage is cheap only if it is well managed.
  - Well managed storage imposes a **retention schedule and disposition instructions** on everything that goes into storage.
  - Otherwise, storage costs grow without limit over time.
- AF strategy: Nothing goes into data storage unless it contains a retention period and disposition instructions.
- In other words, all information assets shall be managed with “records management discipline.”

What is a possible Solution? Information Asset Management as a Strategy

- Under IAM, every “piece” of information – no matter its form or format – is considered an information asset (IA) and is managed according to a single set of principles.
- IAM incorporates and aligns into one discipline the multiple disciplines traditionally associated with Information Management: data management, records management, multimedia management, document management, workflow management, and publications/forms management.

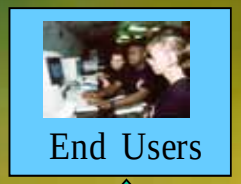
What are the Principles of IAM

1. Information is an asset so long as it has positive value to the enterprise. When it no longer has positive value, dispose of it.
2. Information is a time-related asset; it has positive value for a definable period of time (including, if necessary, in perpetuity).
3. The value of information depends on the ability of the enterprise to discover, access, understand, and consume the information.
4. Each IA shall have assigned to it at the moment of creation a period for which it is to be retained and instructions for disposition of the information at the end of that period.
5. All IA's are to be managed with no intrusion on the end user, maximizing automated information management.

Implementing AF IAM

- **Policy**: All AF data shall be made visible, accessible, and understandable to any potential DoD user as early as possible in life cycle to support mission objectives.
- **Implementing Principle**: AF will invest heavily in metadata to manage all forms of information assets.
  - Every Information Asset shall have associated with it a full set of metadata.
  - AF shall invest in automatic extraction of each IA's metadata. There shall be no intrusion on the end user to create the necessary metadata.

How do we implement the solution?



Functions

-Autogeneration of metadata
-Discovery Engine

-Vocabularies
-LCM Rulesets

-Info Asset Retrieval
-Storage Management

Semantic Layer

Metadata Layer

Infrastructure Services

Technologies

-AMPS/MDE
-Concept Searching

-AF MDE
-CommVault

-CommVault
-Sharepoint

Documents

EMAIL

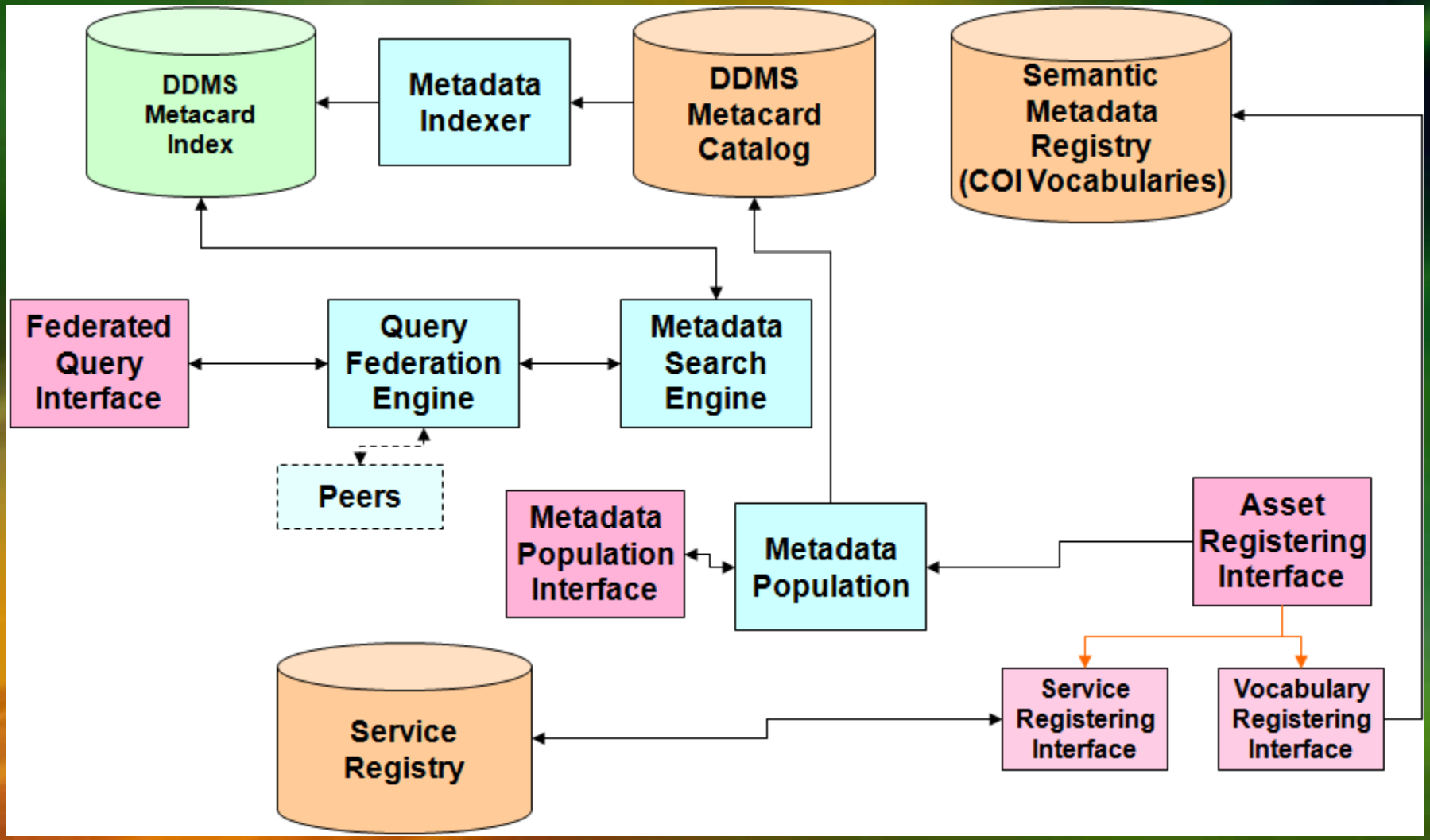
Imagery

Other Information Assets

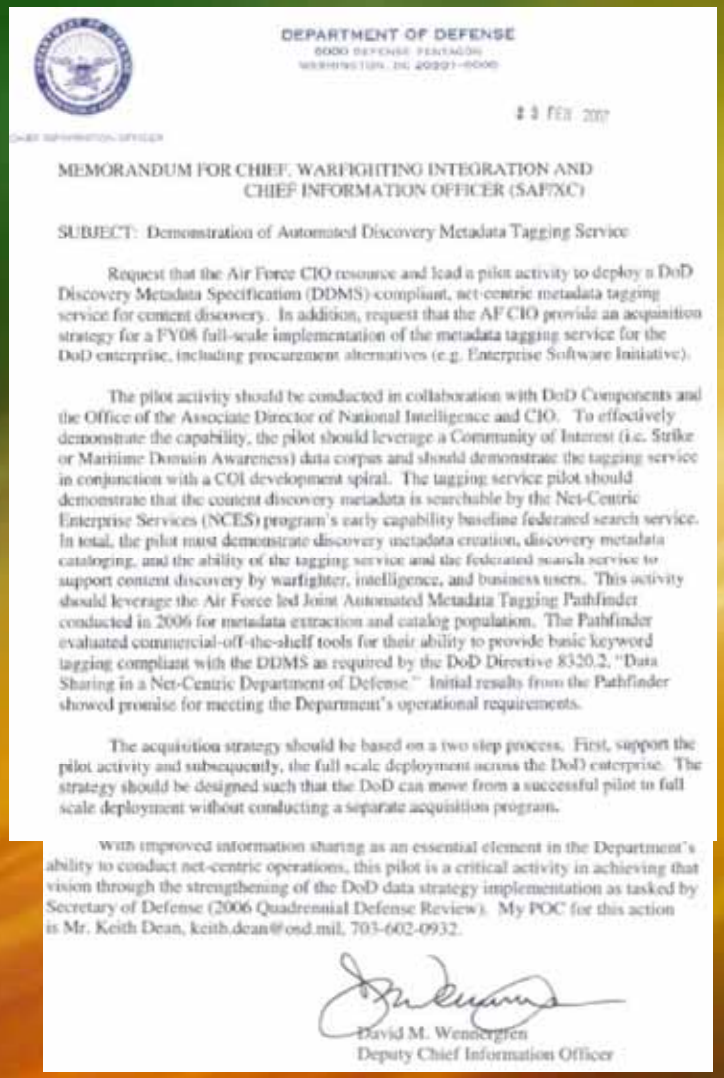
Management and Control of Metadata is accomplished through the Metadata Environment (MDE)

- In IAM –
  - MDE operates as a web service at the server level on all IAs.
  - Any IA in shared space is automatically subject to MDE.
    - Is assigned automatically extracted metadata values – (about 70 metadata elements in AF)
    - Is assigned a retention period and disposition instructions
    - Has its metadata saved in a Metacard & put in Metacard Catalog.

Architecture of the MDE



Metadata Tagging Memorandum



DEPARTMENT OF DEFENSE
6000 DEFENSE PENTAGON
WASHINGTON, DC 20301-6000

3 FEB 2007

MEMORANDUM FOR CHIEF, WARFIGHTING INTEGRATION AND CHIEF INFORMATION OFFICER (SAF/XC)

SUBJECT: Demonstration of Automated Discovery Metadata Tagging Service

Request that the Air Force CIO resource and lead a pilot activity to deploy a DoD Discovery Metadata Specification (DDMS) compliant, net-centric metadata tagging service for content discovery. In addition, request that the AF CIO provide an acquisition strategy for a FY08 full-scale implementation of the metadata tagging service for the DoD enterprise, including procurement alternatives (e.g. Enterprise Software Initiative).

The pilot activity should be conducted in collaboration with DoD Components and the Office of the Associate Director of National Intelligence and CIO. To effectively demonstrate the capability, the pilot should leverage a Community of Interest (i.e. Strike or Maritime Domain Awareness) data corpus and should demonstrate the tagging service in conjunction with a COI development spiral. The tagging service pilot should demonstrate that the content discovery metadata is searchable by the Net-Centric Enterprise Services (NCES) program's early capability baseline federated search service. In total, the pilot must demonstrate discovery metadata creation, discovery metadata cataloging, and the ability of the tagging service and the federated search service to support content discovery by warfighter, intelligence, and business users. This activity should leverage the Air Force led Joint Automated Metadata Tagging Pathfinder conducted in 2006 for metadata extraction and catalog population. The Pathfinder evaluated commercial-off-the-shelf tools for their ability to provide basic keyword tagging compliant with the DDMS as required by the DoD Directive 8320.2, "Data Sharing in a Net-Centric Department of Defense." Initial results from the Pathfinder showed promise for meeting the Department's operational requirements.

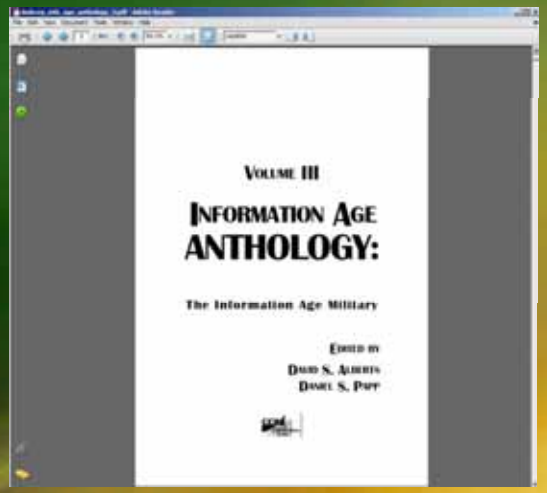
The acquisition strategy should be based on a two step process. First, support the pilot activity and subsequently, the full scale deployment across the DoD enterprise. The strategy should be designed such that the DoD can move from a successful pilot to full scale deployment without conducting a separate acquisition program.

With improved information sharing as an essential element in the Department's ability to conduct net-centric operations, this pilot is a critical activity in achieving that vision through the strengthening of the DoD data strategy implementation as tasked by Secretary of Defense (2006 Quadrennial Defense Review). My POC for this action is Mr. Keith Dean, keith.dean@osd.mil, 703-602-0932.

David M. Wennergren
Deputy Chief Information Officer

- Air Force formed Automated Metadata Population Service (AMPS) Working Group
- NSA formed Information Assurance sub-group
- Participation
  - Government
    - Air Force
    - JFCOM
    - NSA
    - Army
    - DISA
    - Navy
    - DIA
    - NGA
  - Industry
    - Booz Allen Hamilton
    - eCompex
    - MITRE
    - Apache

AMPS Vignette



AMPS

- Automated Metadata Population Service (AMPS)
  - Using vocabularies, generates a MetaCard from an asset automatically
  - MetaCards used for lifecycle management and discovery of data and services

Spiral 1 Scope

- **Produce Discovery Metadata from COI Assets**
- **Exploit Open Standards: UIMA, OWL, WSDL**
- **Label Metacards with CAPCO Markings**
- **Cryptographically Bind Metacards with Original Assets**

- DDMS elements
  - Creator (DISA)
  - Title (DISA)
  - Date (DISA)
  - Subject (AF)
  - Format (Army)
  - Identifier (DISA)
  - Security (NSA)
  - Type (AF)
  - Description (AF)
  - Geospatial (AF/DISA)
- Asset types
  - Microsoft Office
  - PDF
  - Message/email
  - HTML
  - XML
  - XSD, OWL, WSDL
- COI
  - Readiness
  - Blue Force Tracking
  - Information Assurance
  - Generic

Disposition Rules Engine

Individual IA	RETENTION/DISPOSITION "RULES ENGINE"	All Detailed AF Records Schedules
Detailed Metadata Values of IA →	1. *"Rules Engine" matches metadata of individual IA against metadata of ALL AF records schedules using a probability model to find best fit between IA and a records schedule (s).* 2. *"Rules Engine" assigns retention and disposition metadata values from best-fit records schedule(s) to IA. If best-fit meets predetermined criterion, declares IA a record & associates with record(s) schedule.*	← ALL AF records schedules represented as IA's, i.e., the Metacard or "metadata profile" of each records schedule
IA with assigned retention/disposition, record status metadata in Metacard		

Assigning Disposition and Retention

"Rules Engine" –

1. Compares each IA's metadata with metadata of all AF records schedules, finds the best fit(s) and assign retention/disposition from that fit.
2. If fit is "close enough," declares IA a records and associates with records schedule. "Close enough" is AF specification.

- **AF MDE "Rules Engine": designed, not yet completed.**
- Work is "in process" on "Rules Engine" and depends on collaboration between CIO's Office and AF Records Officer.
- **Present Status of IAM**
- Many aspects of IAM are already implemented in multiple AF locations.
- AF intent is to deploy IAM worldwide to 10,000 AF installations and 2 million personnel.

Is IAM applicable to the broader Community?

- IAM is an ***information management infrastructure*** for ALL IM functions & services.
- More broadly, IAM fully solves ***E-Discovery.*** Every IA is discoverable and retrievable.
- IAM fully solves ***Email Records Management.*** Every email sent or received passes through a server where MDE operates; hence, discoverable & retrievable.
- IAM automates all records management decisions previously expected of desktop users.
- In this sense, IAM eliminates the need for desktop user training in RM decision making while greatly improving the accuracy & comprehensiveness of enterprise RM .

Conclusion

Is IAM a worthwhile general model of an infrastructure for managing all enterprise information?

Our Answer: Yes it is.

- We believe the IAM framework will spread widely throughout government & industry.