The "Tagged Data Authority Engine" — Assurance and Data Integrity for Government Agencies and Fusion Centers or of Research

ABSTRACT

At a time when government data centers are closing and consolidating, government agencies also are under pressure to share data across multiple jurisdictions. But this has created a new problem - uncontrolled redundancy and long-term accuracy issues for data as it is shared across multiple agencies and systems. This is especially true at a time when government fusion centers are tasked with gathering data from multiple sources.

Environments which gather data from multiple resources lack the data management tools found within standalone databases. Shared data, once it has left its original system, can be stymied by functions such as version control, rollback, recoverability, merge/purge tools, field lock etc. IDC Government Insights has developed the concept of what it calls the "Tagged Data Authority Engine" (TDAE) to help government agencies enhance data quality assurance by establishing a clear authority path of where each piece of data in a given data set or an XML file comes from – including details on who has authority over that piece of data and where the ultimate authoritative copy of that data resides.

This type of broad cross-agency project can be tackled using existing technologies. But such a project must include high-level coordination across all government agencies, with CIO-level buy-in. A server that is dedicated as a government agency's TDAE is one way to accomplish the concept.

BIOGRAPHY

Shawn P. McCarthy
Research Director
IDC Government Insights

Shawn P. McCarthy is the Research Director at IDC Government Insights. He launched and manages the U.S. Government IT Infrastructure Strategies program, which includes technology recommendations and key cost control proposals for government IT systems. He also issues IDC's semi-annual U.S. Government IT Spending Guides (federal, state and local) and he created the annual U.S. Federal Line of Business Budget Guide, which tracks detailed spending buy federal department, program and more.

Mr. McCarthy is responsible for analyzing primary end-user data and budget data collected both from officially published sources and through surveys of IT managers in government and university offices.

He graduated magna cum laude from The George Washington University with a Masters degree in educational technology leadership. He received a Bachelor of Arts degree in journalism/mass communication from St. Bonaventure University and has a certificate in project management for IT programs. He occasionally teaches graduate-level project management classes and he is the author of two books and also writes the monthly Internaut column for Government computer News magazine.

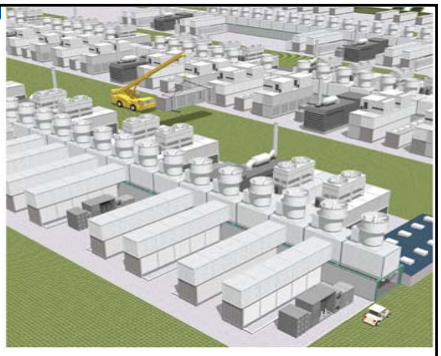
IDC Government Insights

Establishing "Data Authority" For Shared Data Resources

Shawn P. McCarthy
Research Director

July, 2011

How Data Centers Are Changing



- Virtualization
- Networked and remote storage
- Cloud (public vs. private)
- Performance vs security vs service levels
- The future of tablet PCs and gov mobility solutions
  - Needs of the Mobil worker
  - Non-browser based information display
- Temporary/portable/quick install data centers

 Page 2

IDC Government Insights

The Data Assurance Challenge

- Government fusion centers seek to enhance and standardize data received from multiple agencies.
- Information assurance (IA) and data quality assurance are important for government business and national security
- BUT: Information quality can only be assured when there is a clear understanding of
  - Where each piece of data comes from,
  - Who has authority over it (for accuracy, updates, and life-cycle)
  - Where the ultimate authoritative copy of that data resides.

IDC Government Insights

Traditional Database Management Functions Aren't Portable to Large Fusion Centers

- Database management systems offer version control, rollback, recoverability, merge/purge tools, field lock etc.

Some cross-jurisdiction efforts

- To protect data within a specific domain
  - Atomic, consistent, isolated, durable (ACID) properties
- To protect data across multiple domains
  - Data integration tools including data profiling, data quality, and data movement software
- Or - internal network file systems
  - Help with data synchronization and file version control
  - Able to identify and resolve data conflicts
  - Limited to a single file system or set of tightly controlled systems within a single jurisdiction

IDC Government Insights

The X Factors

- Modern data is often shared via …
  - XML files and RSS feeds
  - Exported databases and delimited files
  - Data culled from search engines
  - Documents (MS Word, Excel, PowerPoint)
  - Open source data collections (phone books, credit card transactions, news articles)
- Fusion Centers usually do not have access to the original databases or file systems
  - They receive copies of data
  - Once data is shared this way, it leaves the control of the original owner

IDC Government Insights

Data Integrators Face Unique Challenges

- Increasingly difficult for IT managers to know where their data originated
- Problems with uncontrolled redundancy and long-term accuracy of data
- Tagged data elements help & metadata helps
- BUT - it's still challenging to know
  - How old a data set is
  - How often it should be updated
  - Who is responsible for the ongoing accuracy of each piece of data?

Or…
Who has "authority" over each specific piece of data?

Metadata needs an element called "Tagged Data Authority"

With updates controlled by a "Tagged Data Authority Engine"

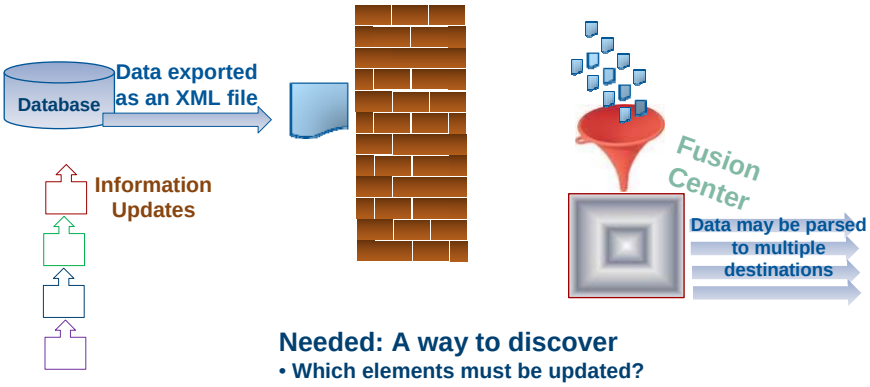
IDC Government Insights

Examples of Establishing "Data Authority"

- Many government forms ask citizens for their Social Security number
  - The only authority for a number is SSA
- During a book discussion, someone might mention a Library of Congress catalog control number
  - The only authority for the number is LOC
- Server IP addresses
  - The only authority for an address is the Internet Corporation for Assigned Names and Numbers (ICANN)
- Every piece of tagged data in a file comes from a specific location
  - Establishing change/update authority for each element enhances data integrity as data is shared

IDC Government Insights

The Limits of Sharing Data Across Organizations

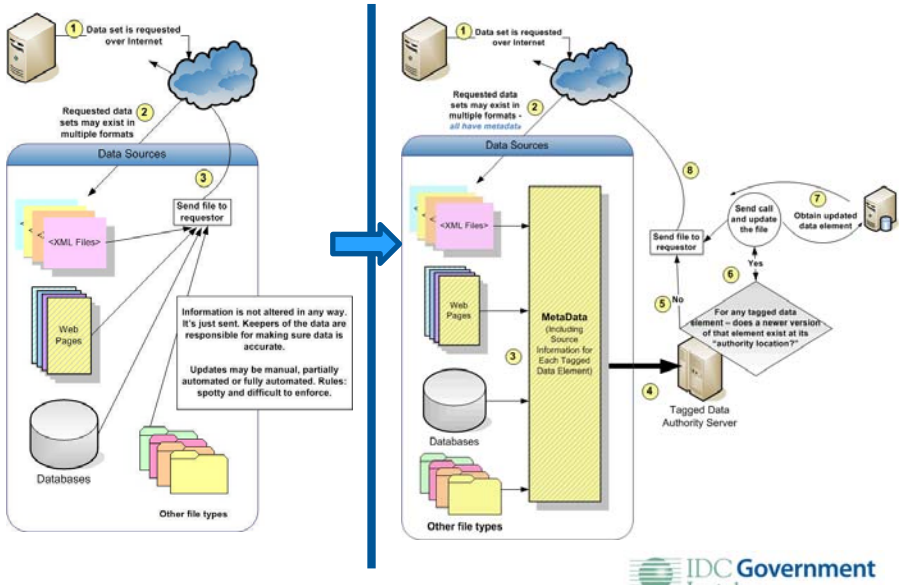


Needed: A way to discover

- Which elements must be updated?
- How often?
- Where does the authoritative copy reside?
- How do we get it?

IDC Government Insights

Basic Structure of a TDAE



IDC Government Insights

Leveraging Metadata

Traditional Metadata

```
<head>
<meta name="description" content="Name and address file" />
<meta name="keywords" content="Citizen name, social security number,
address" />
<meta name="author" content="Joe Jones" />
<meta http-equiv="content-type" content="text file" />
</head>
```

Tagged Data Authority Elements

```
<head>
<Tagged Elements Directory>
<Tag name="SSA_Number" Origin=https://ssa.gov/ss_number/111-22-3333
   Time obtained="06/27/2011 13:02:48">
<Tag name="Citizen Name" Origin=https://DHS.gov/names/search?Michael W. Smith
   Time obtained="02/16/2011 09:31:05">
<Tag name="Citizen Address" Origin=https://IRS.gov/addresses/search?Michael W. Smith
   Time obtained="05/16/2011 16:12:55">
</Tagged Elements Directory>
<head>
```

<Tag Rules>

</Tag Rules>

<Conflict Rules>

</Conflict Rules>

<Merge Rules>

</Merge Rules>

IDC Government Insights

TDAE: Why is it Needed?

<data> </data>

- Current cross-agency data import structures often don't have a way to check the timeliness and accuracy of all tagged data elements.
- Any data element can be tagged, via XML, in a shared file
- Built-in metadata can be included with the file to
  - Establish where the official version of each data element resides
  - Establish rules for data freshness and ownership
  - , then it can be more accurately tracked and relied upon.
- A TDAE promotes data accuracy by calling back to multiple data sources to check for accuracy and updates.
- This type of broad cross-agency project can be tackled using existing technologies, but there must be high-level coordination across all government agencies, with CIO-level buy-in

IDC Government Insights

TDAE Potential

- Could serve as an automated way to perform some types of data cleansing, merging, and updating operations
- Unique because of its mission – reaching across multiple networks and domains, even into domains that are not under an organization's direct control
- Helps update a variety of data sources by monitoring when new sources are available
- Encourages agencies to establish a firm taxonomy for their multiple data sources, including details on where the data comes from, how it is stored, and how copies are checked against an original data source
- Serves as an intermediary server containing a master set of data, capable of tracking where every tagged data element originates

IDC Government Insights

What a TDAE is NOT

- Not a file system or a file mirror
- Not Master Data Management
  - MDM is part of the mix, but the engine is not intended to provide a single reference point for master data or transaction
- Not a distributed parallel fault-tolerant file system
  - It's not about individual files. It's about accuracy of tagged data elements within shared files
- It is not designed for data conflict resolution
  - It is designed to detect such conflicts
  - Resolution can be automated or manual, using other tools
  - Distributed conflict detection - achievable via "version vectors"
    - Mechanism for tracking changes to data in a distributed system with multiple update points
    - Keeps track of update events at different locations, establishing which copy has overwrite privileges.
    - But - few reliable ways to establish version vectors across multiple government agencies

IDC Government Insights

Models Which Might Be Followed For A Government-wide TDAE Design

- The Internet Domain Name System (DNS)
- Akamai Technologies' Content Distribution Model
- Semantic Web Technologies
- The Andrew File System
- Dublin Core
- Coda
- Lustre
- Mark Logic Server
- Ruby on Rails
- Nonschematic DBMS
- Data Integration and Access Software

IDC Government Insights

The Internet Domain Name System (DNS)

- Series of authoritative (root) name servers - which hold all domain name records and associated Internet Protocol (IP) addresses
- It's a distributed database system. Other domain name look "upstream" to the root servers to find latest information for IP lookups
- Local or regional DNS machines
  - Don't copy the full range of DNS info present in the root servers
  - Hold IP information in their cache for a limited time
  - Eventually, they do new lookups to be sure they have the latest official IP address
- Government helped develop the DNS – so it's already familiar with this approach

What's missing

Not designed for other types of files

IDC Government Insights

The Akamai Technologies Approach

Akamai

- A global Internet distributed computing platform - clients are large Internet service providers
- Akamai speeds delivery of content by mirroring it to multiple servers around the globe
- Uses time-stamp files and establish version control for multiple data sets
- Coordinates how continuously updated files are stored on all servers
- Regional servers are supposed to know whether their cache includes the most recent version of a given file
- If a newer version is available the new version is passed to the regions' servers

What's missing

Doesn't break out every tagged data element in a file

IDC Government Insights

Semantic Web Technologies

- Work is coordinated through the WWW Consortium (W3C)
- Seeks to develop ways to share and reuse data across multiple applications, enterprises, and communities
- Approach to linked data is make data reachable and manageable where it sits — via Semantic Web tools
- Provides environment where applications can query data and draw inferences using vocabularies
  - Web Ontology Language (OWL)
  - Simple Knowledge Organization System (SKOS)
  - SPARQL, and others
  - Resource Description Framework (RDF)
    - Provides a metadata data model which has been expanded to provide conceptual descriptions or modeling of information in various Web resources
    - Information can be produced and shared in a variety of syntax formats, (Example - RSS feed, where control of info is retained by the feed owner)

Issue

Based on data meaning, not specifically designed to track updated data fields

IDC Government Insights

Dublin Core

- Describes information online in a way that makes it easier to find
- A Dublin Core Metadata Element Set (DCMES) include 15 metadata elements, including Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, and Rights
- Popular in some fields of library and computer science
- May integrate elements of the Resource Description Framework
- Source and Relation metadata could serve as a partial solution to the data authority problem

What's missing

Additional solution needed to provide lookup and auto-update functions

IDC Government Insights

The Andrew File System

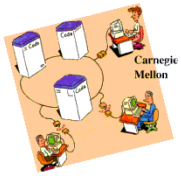
- Distributed networked file system developed at Carnegie Mellon University
- Uses multiple servers while providing a location-transparent file name space
- As with Coda AFS client machines cache files on their local file system, to accelerate subsequent file requests
- Supports access to a limited subset of files in the event of network or server disconnection
- Sets a "pessimistic" file replication strategy. It allows only one read/write server to receive updates. All other servers become read-only replicas

What's missing

Adjustments needed to support a mix of servers – some of which may not want to directly participate in an AFS

Coda

- Distributed network file system, handles caching and version control on a local network, or a WAN (If partners are tightly integrated)
- A client machine reads and writes to a network file system normally -- but while keeping a local copy of all the data it wants to use
- If the network connection disappears, the client's local cache continues to hold the data, logging all updates and reintegrating with the file system when the connection is restored
- Unlike AFS, Coda supports "optimistic" file replication
  - Makes copies of data more widely available
  - This also increases the chances of data conflicts



- Includes multiple manual and automated conflict detection, resolution, and repair tools to help reestablish data authority after a period of disconnection

What's missing

Doesn't necessarily solve data authority issues for data that has "left the building" for integration into external systems

Lustre

- Object-based, distributed file system, popular in large-scale cluster computing
- Open source (but designed, developed and maintained by Oracle Corporation, via its acquisition of Sun Microsystems. Sun acquired Cluster File Systems, Inc., in 2007)
- Via a dedicated metadata server, it can support a single metadata target (MDT) per file system, capable of holding file names, directories, permissions, and file layout
- When it comes to information sharing the four primary design patterns are:
  - One to one
  - One to many
  - Many to many
  - Many to one

Luster or Coda work well for three of the four solutions but not so well for many-to-many systems

Mark Logic Server

- Mark Logic - makes a dedicated XML server & associated XML databases, plus tools to
  - Implement file version control
  - Service-oriented architecture (SOA) rules and policies
  - Some government offices use the Mark Logic server for file management, information tagging, and version control, while using EMC's Documentum as an enterprise content management platform or Microsoft SharePoint modules for integrated collaboration, process management, and document management functions.
- Company has been tapped to
  - Develop the next-generation card catalog for the Library of Congress
  - Create the architecture for the long-term digital archives for the National Archives and Records Administration

Challenge: Still have to work to integrate individual tagged data elements

Nonschematic DBMS

- A relatively new technology area that's sometimes (inaccurately) referred to as a "NoSQL database"
- A variety of products which have some of the characteristics of traditional DBMS, including recoverability and formal query support; but also with advanced search support and the ability to accept data that is labeled or tagged without a schema to define it
- Scalable and typically deployed in cloud environments
- Often used as internal resources for cloud-based services (Amazon's SimpleDB and Google's Bigtable)
- Some derive their approach to data management from simple tagged data programming systems such as MapReduce, Data Spaces, and Memcached

IDC Government Insights

Ruby on Rails (ROR)

- ***Ruby*** is an object-oriented programming language with a syntax similar to Perl plus some elements of Smalltalk and other languages
- ***ROR*** is an open source Web application framework for Ruby that's most often used for the rapid development of applications and integrated systems.
- Mentioned here not because of its server capabilities but because its programming structure -- encourages developers to locate information in a single, unambiguous location
  - Using the ActiveRecord module of Rails to add certain types of inheritance and associations to specific data

Issue

Has been criticized for not being able to scale for very large implementations

IDC Government Insights

TDAE – Likely Solution

- Combination of several of these technologies
- Possibly built on an Andrew File System (AFS) core, but with functionality extended from the file level down to the tagged data element level
- Must be ambitious in its scope (near DNS level) with appropriate level of CIO commitment

 Page 25

IDC Government Insights

Changes Are Needed Before A TDAE Can Be Implemented

- Every piece of tagged content within a file or a database must have a corresponding entry in the metadata information maintained within that file or database.
- Metadata must include specific details on
  - Where the master copy of that data resides
  - Information on how it is time-stamped (or other ways to check to see if the server has the latest version of a tagged element)
- Using a copy of another organization's data? You must know…
  - How to use the metadata to automate checks and to aromatically obtain updated copies
- As shared data is parsed and repurposed, care must be taken to keep appropriate metadata <u>with</u> the data as it is split off from each file

 Page 26

IDC Government Insights

Gov Will Face Standardization Challenges

- Schemas: Should government accept broad industry standards or adopt its own?
- Data control: How should data be updated once it's released across multiple documents and resources (and possibly removed from its XML format)?
- Access: How should a platform be established that can expose all of these different capabilities of available data sets and associated IT services, allowing the government to build applications to suit specific needs?
  - Yes, standard Web services designs using the Simple Object Access Protocol (SOAP) and Web Services Description Language (WSDL) can do this, but extending this capability across multiagency environments can be very challenging.

IDC Government Insights

Standardization Challenges (Cont.)

- Interoperability: Government is dealing with many legacy systems and will for some time to come. This is a challenge unto itself.
  - Eventually information-sharing platforms need to provide controlled vocabularies, data harmonization, and data ownership policies and standards
- Other long-term issues include:
  - Better reuse of prepared information
  - Dynamic publishing across multiple platforms and documents
  - Delivery of content in context
  - Being able to offer contextual information for data, including source, role, situation, mission, geography, and more

IDC Government Insights

Before Initiating

- Standardize security across all data sources
- Implementing MDM is a useful precursor, with a map of all data including data warehouses
- Establish policies and procedures to ensure any work done for the TDAE is consistent with the enterprise data governance policies and procedures
- A data integration project that will create an enterprise-wide integrated data environment
- A plan to coordinate database operations, including scheduled downtime for software and server maintenance and upgrades, backups, migrations, etc.

IDC Government Insights

Conclusion

- The TDAE is an idea whose time has come.
- But a TDAE-style solution cannot be constructed without broad buy-in at multiple levels of government, and across multiple agencies
- Thus the solution can't be approached haphazardly. It first needs a champion at the highest levels of government IT management
- Federal or statewide CIO councils are the likely starting point. If agreement is reached there, committees then need to be formed to address how metadata elements can be created
- Then, the idea becomes a project management issue, with associated technologies, deliverables, and internal marketing efforts that can be used to get buy-in across all departments

IDC Government Insights

Questions?

Shawn P. McCarthy

Research Director

smccarthy@idc.com

IDC Government Insights