AUTHORITATIVE DATA SOURCE (ADS) FRAMEWORK AND ADS MATURITY MODEL[1]

(Practice-oriented Paper)

Frank J. Ponzio, Jr.
Symbolic Systems, Inc., New Providence, New Jersey
fponzio@symbolic.com

Abstract: Throughout the Federal Government, including the Department of Defense (DoD), there is a pervasive need to have current, reliable, and trusted data from what are termed Authoritative Data Sources (ADSs). This requirement has grown increasingly important as the military transforms. The transformation requires individual organizations, each using an array of independently developed stovepipe systems, to function on the battlefield with other weapons systems, services, and friendly nations, sharing communications networks and data. To ensure the accuracy of the data being provided by ADSs, new methodologies and metrics must be initiated. This paper proposes adoption of an Authoritative Data Source (ADS) Framework to analyze and adjudicate information quality issues prior to publishing data for consumer use and an ADS Maturity Model to rate data providers.

Key Words: Information Quality, Data Quality, Authoritative Data Source, Data Quality Feedback, Adjudication, Maturity Model, Framework

INTRODUCTION

Many communities of interest, within and outside of the Federal Government, rely on ADSs for certain types of data. An ADS's product can range from simple lists of codes and associated names to complex work products like architectures. Within the Department of Defense (DoD), for example, the areas of systems architecture, Command and Control (C2), and Situational and Battlefield Awareness, all need reliable, trusted data from ADSs for mission success. This need is heightened by the shift to net-centric operations within DoD.

This paper discusses the adoption of an Authoritative Data Source (ADS) Framework [4] in organizations that rely on external sources for data or that distribute their data to others. The author proposes that this framework be adopted to define a repeatable process for improving the quality of ADS data products. The paper also proposes that an ADS Maturity Model [4] be adopted to assess the quality and risks associated with a specific ADS product. A maturity model for data would provide a standard against which data sources could be assessed, similar to the Capability Maturity Model® for Software, which was broadly accepted as the de facto standard for assessing and improving software processes [2], [9].

Adopting these two models is a transformational, scalable solution for the ADS community. For ADS providers, it is a transition plan to ensure the high quality of data that is needed by all data consumer communities. For ADS product users, this solution offers a means to assess confidence in the quality of

the data and determine the risks associated with the ADS product they are using. They thereby become more knowledgeable users.

This paper, in addition to proposing the ADS Framework and ADS Maturity Model, also presents some considerations, proof of concept experiences, lessons learned, challenges, and conclusions that are associated with this initiative. It does not detail all of the tasks (programs, checklists, reports, etc.) required to implement these models.

Since the ADS Framework and ADS Maturity Model have recently been adopted by some organizations within the U.S. Army, this paper uses the military in its examples. The use of these models can be applied as a standard data quality improvement process for any government or commercial organization that relies on external source data for successful operation or that wishes to confirm the quality of the data that it disseminates to others.

BACKGROUND

Organizations are increasingly relying on systems to conduct their business. These systems typically interoperate with other systems, both internal and external to the organization. As this reliance on systems has grown, so has the reliance on data that is provided by others. This data may be used or published as is or may be integrated and manipulated. The number of inter-system transactions and information exchanges has proliferated even more with the expansion of Internet use. Often the data provided by a source information provider is critical to the successful operation of the receiving organization.

For example, the Department of Defense is transforming the military from individual service arms, each of which develops its own battlefield systems for its own use, to a more collaborative, inter-networked force. Not only are the digital systems within any one service now required to interoperate and share data among themselves, but also many systems developed by and for individual services are part of the joint battlefield's Tactical Internet. Because digital systems were initially developed as independent stovepipe systems, their underlying databases and systems requirements are not standardized. In addition, the networking information they require to communicate with each other across the Tactical Internet also differs from system to system. Consequently, different battlefield systems rely on different, and in some instances, many Authoritative Data Sources (ADSs) to provide key information for the successful interoperability of their systems. An overview of how one organization in the U.S. Army has applied the ADS Framework and ADS Maturity Model and lessons learned from this initial proof of concept is covered later in this paper.

To date, ADSs could be considered a "cottage industry," where many ADSs are providing a variety of data products using a multitude of methods with multiple risks regarding the data they are providing. The users of data from ADS products face risks when using this data. The following types of questions should be asked to help mitigate some risks:

- Am I using the same version of this data that everyone that I need to interoperate with is using?
- Should I all be using a later version? Does a later version exist?
- Have I properly taken into account changes between versions of the ADS's data?
- Is one ADS's data consistent with the same data from other ADS?

The problem facing most organizations is what, how, and when to address the risks. In a best-case scenario, everyone is taking the necessary actions to address these risks. However, this results in significant duplication of activities across organizations, probably with varying results. In a worst-case

scenario, none of the risks are being addressed and wrong information is being used for system tests. Unfortunately, everything "looks good" until ADS issues surface in integration tests, exercises, or when the data is required in a production environment.

At the same time, many organizations, in addition to being data consumers, are also data providers. Just as they wish to protect themselves from risks associated with the data they consume, organizations may also wish to be viewed by their customers as providers of high-quality data. Providing products of poor data quality is costly. In addition to the potential for lawsuits, there are other associated costs (i.e., operational, lost business opportunity, and unnecessary expense associated with disseminating flawed data [3].) Additionally, the necessity to provide high quality data is in some cases, for example in the Federal Government with the passage of the Data Quality Act, even a legal requirement [5]. ADSs can adopt these models within their own organizations, prior to data publication. Adopting a framework through which data is reviewed prior to distribution can limit the provider's liability, in addition to enhancing its customer relations.

MITIGATING RISK AND PUBLISHING BETTER DATA PRODUCTS THROUGH PROCESS CHANGES

Organizations that rely on external information can integrate new processes into their operating procedures to help mitigate both the risks of importing and of disseminating bad data. This involves setting up a framework under which data will be reviewed and data issues will be adjudicated with the source provider prior to its use or the data consumer prior to its dissemination. It also incorporates rating the maturity of the data based on the analysis and other efforts made by the source to confirm its quality. Consistent use of these new processes being incorporated into a data quality standard operating procedure will encourage ADSs to make every effort to provide a better quality product to their data consumers.

ADS Framework Model

The proposed ADS Framework Model is presented in Figure 1.

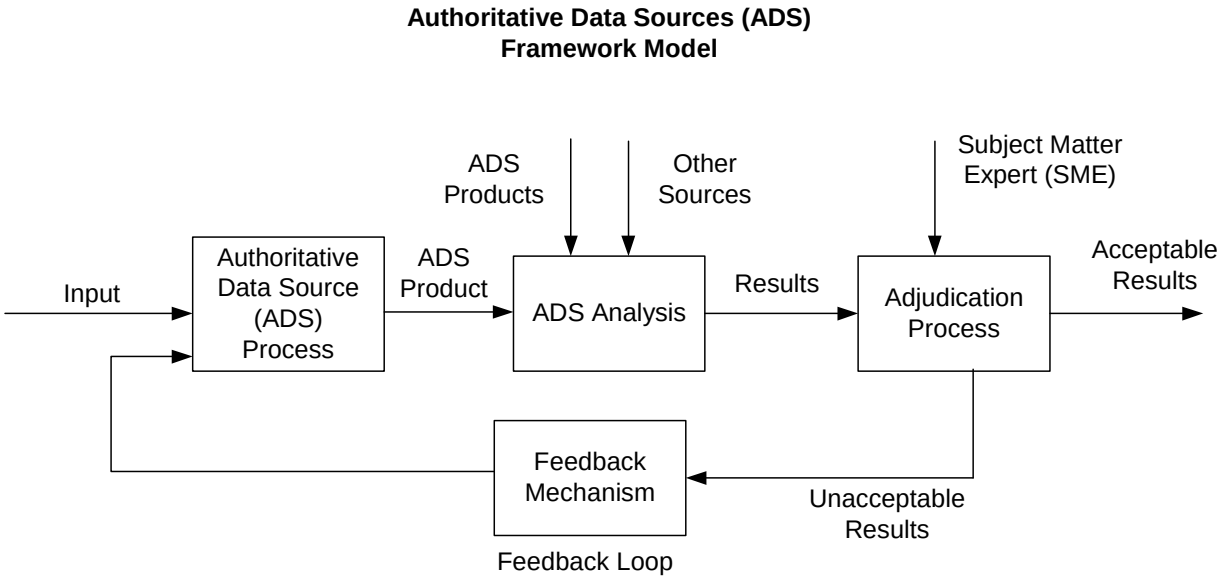

Figure 1: *Authoritative Data Sources (ADS) Framework Model*

The ADS process receives input that an addition, modification, or deletion is required in an ADS product. This change could be as simple as a change to a code list or as complex as a change to a multi-node network architecture that is used by others in their development plans.

The updated ADS product is then submitted and subjected to an ADS analysis, the intent of which is to improve and validate the quality of the ADS product and to reduce the risks identified above for users of the ADS product. The ADS analysis can use prior versions of the ADS product and, if available, data from other sources for the ADS analysis. The ADS analysis might include a comparison and contrast analysis using the prior version and a uniformity and consistency analysis using data from other sources. The results of the ADS analysis are then submitted to an adjudication process. The adjudication process would be performed using the developers of the ADS product and Subject Matter Experts (SMEs), who are generally users of the data, possibly in the form of a data review board, to scrub the results. Adjudication would ascertain if, based on the input that was provided to the ADS process, the ADS results are acceptable or unacceptable. The unacceptable results would be used as feedback to the ADS process. This feedback loop process would continue until only acceptable results are achieved.

At this point, the ADS product and the ADS analysis results would be published and appropriate alerts distributed to users.

The use of the ADS Framework model is transformational for both the ADS provider and user. For the ADS provider, the ADS analysis negates the resource burden of developing and publishing an analysis. The adjudication process adds additional expertise to the process and expands the sphere of knowledge associated with the ADS product. For the user, who relies on the ADS product, this is the equivalent of an Underwriters Lab (UL) [6] approval of the results with full transparency and disclosure of the ADS product and ADS analysis.

Another part of the ADS Framework is to have each ADS provide an ADS analysis with each version of the product that contains the following information:

- The details of the additions, changes, and deletions between this version and the prior version.
- The types of internal quality assurance validations that have been performed on the product. This would include duplication, consistency, uniformity, etc., types of checks.
- The location of other sources of which they are aware, who provide the same or similar information. Users could decide if, when, and how to use this information as part of their risk mitigation plans.

This information is captured in tag information associated with each ADS product. This disclosure to potential users would be used as part of the ADS risk assessment.[2]

Expanding the ADS Framework Model

The framework is scalable as a part of a semantic heterogeneity[3] process. This multi-tiered approach is helpful to resolve questions, conflicts, or issues that arise when there are multiple data source providers. When multiple sources supply shared data elements the data product characteristics in each source and in the receiving application are cross referenced to the common data architecture. A crosswalk analysis matrix, also referred to as a data feed table[1], is developed. For example, if two sources provide addressing information, the following types of analyses are performed.

[2] It is important to understand that publication of ADS data is only a snapshot in time. In all likelihood, the updated additions, changes, and deletions were in effect before the publication was issued.
[3] The identification of semantically or conceptually related objects in different databases.[6]

- A crosswalk analysis between data products to determine data elements that are shared, and if found, any naming conflicts and attribute differences. For example, if one source uses the naming convention SYSTEM_ID and another uses ID_SYSTEM, these may be referring to the same data element and require mapping. Furthermore, if one source provides addressing information that is incomplete without additional addressing information from the other source, inter-product analysis may be required when one source provides updated data.
- For common data elements with different naming conventions or field attributes, a set of mapping tables for standardizing names and field attributes prior to import into the recipient database.
- When data is published, a comparison of the data that is shared between sources in order to highlight any differences or inconsistencies between data sources or conflicts in dependencies between related data elements from data sources.

Figure 2 expands on the ADS Framework Model shown in Figure 1, to show how the model can be expanded.

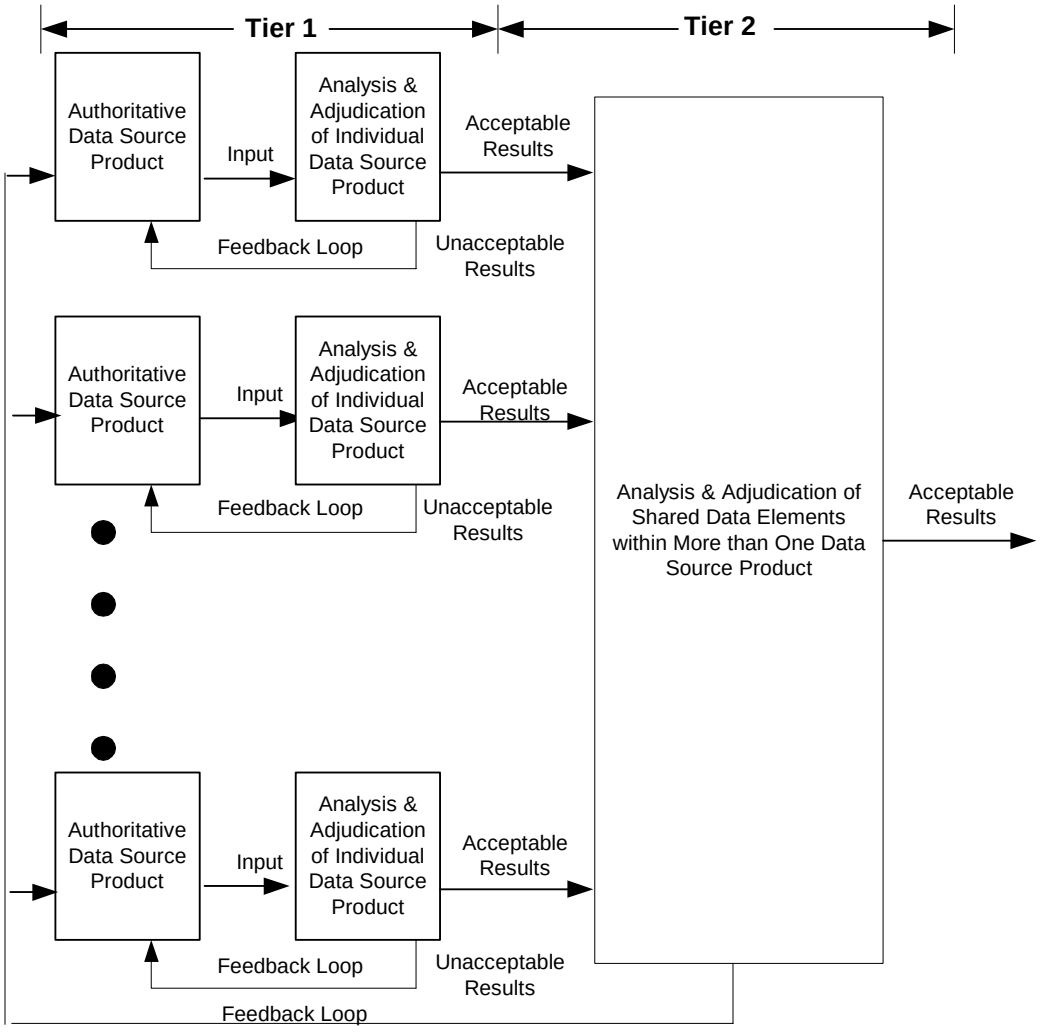

Figure 2: *Authoritative Data Sources (ADS) Framework Model scaled to address multiple data sources.*

Implicit in a two-tier framework, as in the figure above, is a Community of Interest (COI). The framework now involves, at a minimum, three and potentially more stakeholders: the data recipient and the data sources whose products were included in the crosswalk analysis. The resolutions of any data issues are resolved as a community, since any or all of the data sources may be affected.

ADS Maturity Model (AMM)

The use of a maturity model is beneficial for everyone—ADS providers and users. It provides an implied transition plan for each ADS to strive to improve its products. It provides users with the knowledge of what level of maturity is associated with the ADS product. In situations where there are multiple sources for the same or equivalent ADS product, users could use the maturity model level designation of the source to determine which source to use and/or in which source they have the most trust.

The maturity model has levels that would indicate the risk management steps that were addressed by the ADS provider. For example, if there is a five-level model, the maturity levels might be determined as follows:

Maturity Level	**Risk Management Steps Taken**	**Information Source**
0	No ADS analysis is provided.	
1	The adds, changes, and deletion between successive versions are provided and approved by the provider.	Provided in the ADS analysis
2	Duplication, consistency, and uniformity checks were performed.	Provided in the ADS analysis
3	The ADS results have been scrubbed by at least one SME.	Provided in the adjudication process
4	Multiple SMEs and multiple users have accepted the ADS analysis results	Provided in the adjudication process

Table 1: *The ADS Maturity Model*

The AMM level also captures tag information to be used by users as part of the ADS risk assessment.

Example

Because of the broad and in depth interest in the use of architectures within the Federal Government, an architecture example is used below as an example to assist readers in determining the applicability of adopting this ADS Framework to their data sources.

The organization producing a data-based architecture would be considered an ADS to the organization using that architecture in its systems or applications to build a product. An architecture may consist of multiple components that typically present various “views” of the architecture. For example, in the DoD Architecture Framework there are multiple operational, system, and technical views of the architecture[8].

A change to an architecture can affect the network structure, which impacts battlefield communications. Therefore, adjudication of any changes by all stakeholders is essential to ensure that all affected parties have had input and are aware of the impact. In addition, a framework is also essential to ensure that a change in any one view is reflected in all other published views.

For example, if an existing radio were replaced with a radio that uses a new technology and provides additional capabilities, this could have broad impact on the architecture. The group creating the architecture would develop a new architecture version reflecting the new radio and the necessary

connectivity changes in the various views. The ADS analysis would detect all of the changes made in all of the views, comparing them to the prior version of the architecture and other authoritative sources. The stakeholders, to ensure that all changes required for the new radio have been reflected in the new architecture version, would submit these ADS results for adjudication. In addition, they would ensure that the desired communication functionality will be achieved by the warfighter. If there were multiple SMEs and users involved in the adjudication process, the organization that developed the new architecture (the ADS) would be at a maturity level 4.

CONCEPT VALIDATION

Our company has been heavily involved in the Army's efforts to produce initialization data for the digitized weapons systems now being deployed. As described earlier in this paper, these systems rely on accurate addressing and networking information for routing communications and C2 information via a Tactical Internet. The network and addressing information is provided by a number of sources within the Army and DoD to the group in charge of providing the initialization data loads. In order to produce the data loads for the various digitized battlefield systems, the information is imported into a database, manipulated and enhanced, and then exported in a number of different file formats with different field and attribute requirements to meet the unique requirements of the individual weapons system.

Figure 3 shows how the ADS Framework is applied at this site on both the input and output sides of the system. Data products received from ADSs are analyzed and adjudicated before they are integrated into the database. Data products produced from the database are analyzed, posted for review, and adjudicated by a Data Review Board before being published. Action items that are the outcome of the adjudication process are assigned to the original ADS, the initialization data product developer, or the ultimate consumer, as applicable.

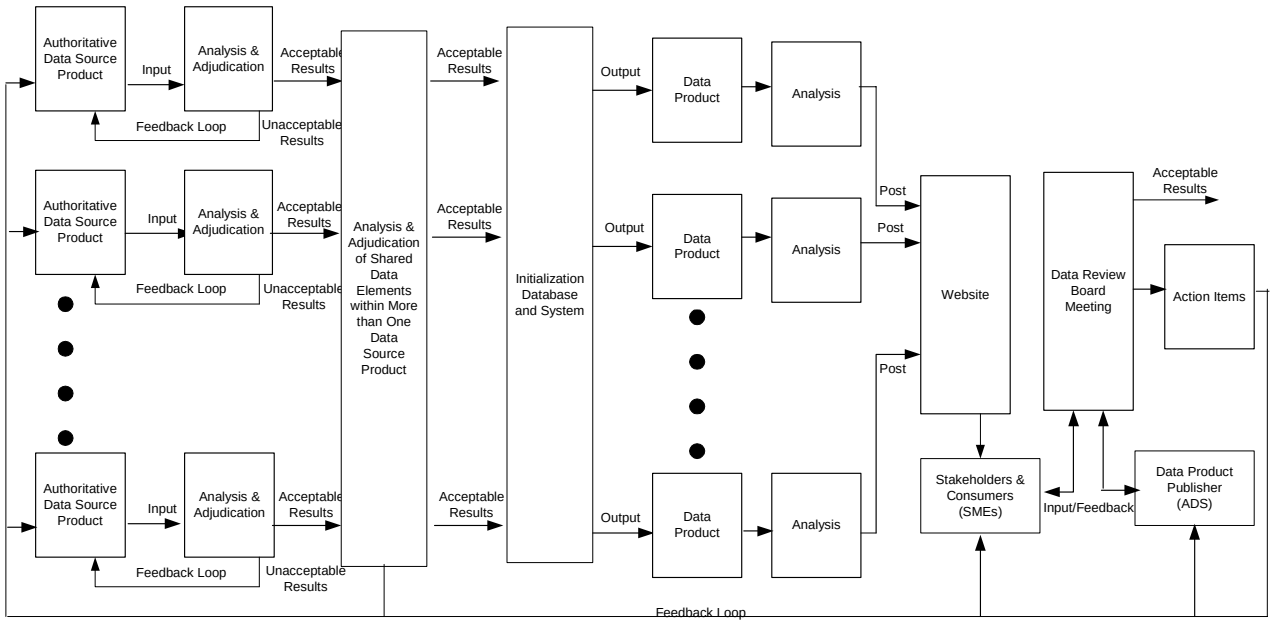

Figure 3: *ADS Framework applied both on the input and the output sides of the system.*

Applying the ADS Framework to ADS Data

As leads on the Data Harmonization team, we discovered early in our engagement that data provided by ADSs was not necessarily accurate or complete. In addition, we discovered that some data elements common to more than one ADS product were conflicting. Some sources, in order to meet their deadlines,

provided what they had at the time it was required without doing any quality checks. Others made undisclosed data changes between versions that affected other systems and programs.

The Data Harmonization team developed a number of analyses, specific to the ADS providing the data, in order to ensure that data is accurate before importing it into the initialization database. When new source data was provided, analysis reports were produced and a feedback/correction cycle continued until the provided data was accepted. Hence, an informal ADS Framework was in operation. This informal mechanism led to formalizing the process by developing and publishing to all ADSs and stakeholders the ADS Framework process that would be used to analyze and adjudicate issues related to the source data upon which we relied.

Applying the ADS Framework to Data Products before Publication

Our customer, being a data product publisher as well as a data consumer, was anxious to provide each systems developer accurate data load information. The goal was to reduce the number of data-caused errors when the systems and networks were tested for interoperability on the test floor and to minimize any communications errors caused by incorrect network or communications addressing data on the battlefield. In the former case, both time and money are wasted if data error and correction cycles extend the testing cycle; in the latter case data accuracy can determine whether a soldier can communicate or a field commander receives complete and correct information for making battlefield decisions. Hence, data errors can cost lives. The customer's goal was to be a maturity level 4 data provider. To this end, we applied the ADS Framework to the developed data products through a formal Data Review Board.

The board, which meets whenever there is a new data load release, reviews with the battlefield systems developers the data that will be provided for their systems and reviews with the operations group from the test floor the proposed network. Prior to a Data Review Board meeting, the data to be incorporated in the final files are analyzed and views of the file data are developed for each consumer. All stakeholders, including the system developer SMEs, are invited to participate either in person or by teleconference. The analysis provided and SME participation in the process rates the data provided in the data loads at a maturity level of 4. The success of the Data Review Board with its adoption of this quality process has virtually eliminated data-related error reports from interoperability testing.

Developing a Data Product Integration Architecture

In order to help determine the analysis requirements for data products (both provided by ADSs and published by our customer), we developed a data product integration plan based on the DoD Architecture Framework[8]. However, because the DoDAF was primarily developed to document systems, not data, additional views expanding the details of how data elements flow through the process as well as map to and interact with other data elements were added. Descriptions of the key systems views are described below.

Data Exchange Matrix

This system view documents information about the data/information we exchange with ADSs. The communications paths included in the matrix followed the process of receiving and of feeding back information. In most cases, the source system element was specified as a particular application or database, but in other cases the system was identified as a telephone. (This is used in some cases as the a feedback mechanism for issues about a received data file.)

Developing this matrix proved helpful for analyzing whether there were any inefficiencies in the methods used to exchange information and data files with our ADSs.

This view captures, in spreadsheet format, the following information:

Column	Description
Operational Source Node	The organization from which the data is received, for example the ADS that sends us data products or our feedback.
Source System Element	Information about the system that is providing the information, including as applicable, the hardware and application source of the data.
Content	What is exchanged, generally in business terms.
Receiving System Element	Like the source element, what system or application receives the information.
Receiving System Node	The organization, or in our case, the specific group that receives the data.
Media Format	The file format of sent and received information, for example a flat file, .sql file, .xls file, etc.
Format	The field-level format, for example, pipe delimited or database.
Protocol	The protocol for receiving the data. Protocols varied widely—from via email attachments to via SQL scripts.

Table 2: *Data Exchange Matrix*

Inter-Product Crosswalk Analysis Matrix

This matrix, which was an extension to the standard DoDAF, depicts the common data elements and the associated fields that are used to identify the equivalent data element in other data products. In the example in Table 3, three different databases refer to the same data by three different field names, i.e., NETMASK, SUBNET_MASK, and SUBNETMASK. Determining and documenting that the same data element is stored under different names in different source and recipient databases was used as a basis for analysis of instances of an element.

ID	COMMON ELEMENT	FIELD TYPE	FIELD NAME	SOURCE SYSTEM[4]
278	NETMASK	VARCHAR(2)	NETMASK	Initialization Database
279	NETMASK		SUBNET_MASK	ADS 1
281	NETMASK		SUBNETMASK	ADS 2

Table 3: *Sample of Inter-Product Crosswalk Analysis Matrix*

Common Data Elements Matrix

This spreadsheet matrix, a variation of the crosswalk described above, specifies only those elements where the field type attributes or field names differ. In addition to highlighting the data elements that are used in more than one system, it identifies the data asset in which they are stored. For example, the field label PLATFORM_ID is in the initialization database under the name PLATFORM_ID, and appears in one source's data product also under the name PLATFORM_ID, but in another source's data product under the name NODE_ID. This matrix provided a basis for analysis of data through the initialization process. Columns were: Item Name, Source System, Data Element Name, plus a reference number to other developed matrices.

[4] The actual names of the source systems have been removed from this sample.

Data Product Integration Mapping Matrix
This matrix pairs the crosswalk data elements that are required to be mapped. It broke down the crosswalk information in the Common Elements Matrix to a greater level of granularity for specifically mapping the table/field name in one source to another source. For clarification, Table 4 uses the same data element (SUBNET) and sources as shown in Table 3 above.

ID	SOURCE NODE A[5]	INFO/OP ELEMENT	SUB INFO ELEMENT NAME (Table Name)	ITEM NAME (Fields)	ID	SOURCE NODE B[5]	INFO/OP ELEMENT	SUB INFO ELEMENT NAME (Table Name)	ITEM NAME (Fields)
173	Initialization Database	ACCESS MDB/PIPE DELIMITED FILES (DATALOAD)	SUBNET	NETMASK	173	ADS 1	.TAR/.GZ FILE (DATALOAD)	BLOCK_IP_R	SUBNET_MASK
174	Initialization Database				174	ADS 1	.TAR/.GZ FILE (DATALOAD)	UNIT_HOST	SUBNET_MASK
25	Initialization Database	ACCESS MDB/PIPE DELIMITED FILES	SUBNET	NETMASK	25	ADS 2	PDF/.TAR/.GZ FILE (DATALOAD)	SUBNET	SUBNETMASK

Table 4: ***Sample from Data Product Integration Mapping Matrix***

Lessons Learned

The application of an ADS Framework process has been an overall success, but it has also pointed out through its use in a data-critical environment that there are a number of lessons learned that should be applied when taking this approach to receiving and providing data products.

Make the ADS Framework part of the culture
The ADS Framework must be consistently used and accepted by the stakeholders involved on both the data provider and data consumer sides. When it is inconsistently applied, data quality will vary.

Sell the concept to ADS management
Although technical personnel tend to be the community that is most involved with the day-to-day issues related to data, it is important to garner management support for the process. Early in the adoption of the ADS Framework model we realized that the ADS managers, who were stakeholders in improving the quality of the data they provided, passed that attitude down to their staffs. This resulted in less finger pointing, better cooperation, and faster turn-around of corrected data products.

Limit review and feedback to the immediate data supplier or consumer
Even though we would have preferred to take a more holistic approach and have all SMEs review all of the data as a whole, most had limited interest in data or data issues outside of their areas. Therefore, although participation in the Data Review Board was open to all stakeholders, we only required SME participation in data reviews that affected their particular battlefield system. In addition, we created data views specific to their systems.

Provide feedback reports to data suppliers and feed forward reports to consumers
No one wants to display their dirty laundry or their dirty data. We create analysis reports for the supplier that detail all of the data problems and provide a separate set of reports for the consumers that only highlight data issues that affect them. For example, we provide data suppliers with reports that show

[5] The actual names of the source systems have been removed from this sample.

inconsistencies, missing records, duplicate records, errors, etc. Once the ADS provide an acceptable product, we provide the consumer with a separate set of higher-level feed forward reports, for example, delta reports that specify data changes since the previous version.

Distribute feedback in an objective, structured, useable electronic format
Although feedback is often verbal, an objective analysis that quickly points out the data issues you're describing should be provided. We found that providing reports to ADSs in an Excel format not only was easy for them to use, but also provided the ADS with a format they could manipulate or import into other analysis programs.

CHALLENGES

The challenges that need to be addressed to accomplish this proposed ADS transformation are:

- Championing of this initiative by a single or multiple organizations.
- The promulgation of the ADS Framework and the associated AMM to the various communities, for example, Architecture and C2.
- The "deployment" of the ADS Framework to the ADS community.
- The oversight and monitoring of adherence to the ADS Framework.
- Reducing the labor-intensive aspects of the process.
- Securing funding for this initiative.

CONCLUSION

The introduction of the ADS Framework concepts and the AMM for ADS is transformational for following reasons:

- It recognizes that the current situations are not going to be able to support net-centric operations and that improvement is necessary.
- It provides a framework for transition.
- It provides measurable maturity goals to accomplish the transition and therefore achieve the transformation.

It is also scalable, in that the model is as relevant for an ADS product that provides code lists as it is for an ADS product that provides a complex architecture.

It is transformational in that it expands on the concept of metadata tags by adding data quality information to the current administrative information. At the same time, for this additional effort, it provides ADS providers with the incentive to provide high quality data and ADS users with needed risk assessment information about the data on which they must rely.

REFERENCES

[1] Brackett, Michael H., *Data Sharing – Using a Common Data Architecture.* John Wiley & Sons, Inc. New York, NY. 1994. p. 248

[2] Herbsleb, James, Carleton, Anita, Rozum, James, Siegel, Jane, Zubrow, David, *Benefits of CMM-Based Software Process Improvement: Initial Results.* Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA. August 1994

[3] Kim, Won, Choi, Byoung, "Towards Quantifying Data Quality Costs." *Journal of Object Technology*, vol. 2.4. July-August 2003. pp. 69-70

[4] Ponzio, Frank J., *Authoritative Data Source (ADS) Framework*, Symbolic Systems, Inc. White Paper. November 5, 2003

[5] Treasury and General Government Appropriations Act 200, Public Law 106-554; H.R. 5658, (generally referred to as the Data Quality Act) can be found on the web at http://www.nrc.gov/public-involve/info-quality/pl106-554.pdf. (Data last accessed 23 August 2004.)

[6] Underwriters Laboratories Inc. (UL) is an independent, not-for-profit product-safety testing and certification organization. The company's website is http://www.ul.com. (Data last accessed 23 August 2004.)

[7] Wong, Brian T., *The Abstract Data Interface.* Master's thesis. Massachusetts Institute of Technology. 2001

[8] www.dod.mil/comptroller/bmmp/pages/arch_arch_home.html provides an overview of the DoD Architecture Framework, the various views that comprise the architecture, and other architecture policies and initiatives within the Department of Defense. (Data last accessed 23 August 2004.)

[9] www.sei.cmu.edu/cmm. (Data last accessed 23 August 2004.)