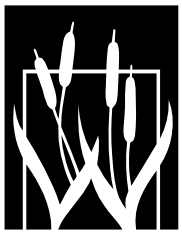

Whitemarsh
Information Systems Corporation

Analysis

OF

(draft) DoD 8320.1-M-1, Data Standardization Procedures
(dated)19 November, 1996
The Proposed Draft
DISA Approach to DoD Data Standardization

Note to the Reader: The analysis contained herein was based on the 19 November, 1996 version of "the 8320". A revised version of that document, dated 28 February, 1998 has been reviewed. This revised document occurred almost six months after this analysis was presented and thoroughly discussed with key members of DISA's Data Standardization team. The changes contained in the 28 February, 1998 version of "the 8320"document do not in anyway reduce the severity of the findings contained within this analysis. The severe problems remain.

Whitemarsh Information Systems Corporation
2008 Althea Lane
Bowie, Maryland 20716
Tele: 301-249-1142
Email: mmgorman@wiscorp.com
Web: www.wiscorp.com

Table of Contents

EXECUTIVE SUMMARY

The current draft of the DoD 8320.1-M-1 Data Standardization Procedures manual and its attendant software system are sure to be changed and supplanted by an adequate DoD data standardization procedure in the near future. This change will occur for the following reasons:

- Federal standards policy requirements. The DoD is required to follow public law (PL 104-106 and PL 104-113 (e.g., OMB Circular A-119 now public law)) that requires federal agency use of standards developed by private consensus standards organizations at the national and international level (e.g. ANSI, ISO, and IEC). This means in practice that the DoD (Draft) Data Standardization Procedures must give way to the ISO standard 11179 on the same subject.

- Inadequacy for its intended purpose. The current DISA draft data standardization procedure cannot achieve the objectives of DoD data standardization because of a critical conceptual error in the DISA approach to data standardization. This critical conceptual error is its inappropriate definition of a data element. A data element in DISA's approach is defined at the level of granularity of a column within a table. For sharing data between different database systems, however, a data element must exist in a one-to-many relationship with columns within tables (or any other type of container such as file, screen, process, report, etc.). Because of this critical and methodologically fatal error, the DISA data standardization approach cannot ever achieve the objectives and goals of DoD data standardization.

- Lack of scalability and extensibility. The current draft DISA data standardization approach and procedure addresses only one of the five classes of databases within an enterprise's total data architecture. This current procedure does not allow DoD to share data across the full range of databases that currently exist and that are under development for original data capture (also known as OLTP database applications), TDSA (transaction-data staging area) database applications, subject area databases (also known as ODS (operational data store) database applications), wholesale and retail data warehouses, and finally, reference data database applications. The DISA model addresses only the first class of data architecture: original data capture (OLTP).

1.0 Overview

The following analysis of the DoD Data Standardization Procedures document -- DoD 8320.1-M-1 has been reviewed by others who concur in its findings and conclusions. The purpose of these comments is to support DoD data standardization progress through procedures and methodologies that:

- Are consistent with federal government standards policies,
- Are based on sound semantic metadata and data engineering principles,
- Have been successful in industry and government agency contexts, and
- That do not suffer from the problems identified in the DoD Data Standardization document, DoD 8320.1-M-1.

2.0 Background: DoD and Data Standardization

The U.S. Department of Defense, through its Defense Information Systems Agency (DISA), has promulgated several documents and software systems related to data standardization. The document with the greatest impact on database and data administration is the (Draft) Data Standardization Procedures, DoD 8320.1-M-1, November 19, 1996. This document sets out the procedures DoD agencies and contractors are mandated to use to develop standard data. Supplementing this manual is the DISA software system Defense Data Dictionary System (DDDS).

Data standardization approaches had already been deployed in various information technology organizations within DoD agencies prior to the establishment of the Defense Information Systems Agency (DISA). The formulation of the current draft procedure for data standardization did not however, synthesize the most advanced approaches that had been developed in these predecessor projects.

The critical conceptual error in the current DISA draft procedure document is found in its definition of the "data element." This error appears to have resulted from a misinterpretation of a Federal Information Processing Standard (FIPS) on data standardization published in the mid 1970s. This FIPS states that a data element is to be the semantic metadata representation for fields within files.

The purpose of this FIPS was to standardize the metadata and the names of fields within COBOL files by connecting--in sequence–1) a prime word to the file's name, 2) with a string of modifier

names to characterize the field's meaning within the COBOL record, and a finally 3) a class word to characterize the field's purpose or data type. This FIPS guidance, mistakenly considered relevant for flat-files of the 1970s, was apparently also mistakenly considered adequate for the evolved relational and multi-dimensional file structures of the 1990s because this three part data element construct formed the basis for DISA's data standardization guidance for implementing the DoD 8320 Policy.

The current 8320.1-M-1 DoD data standardization strategy must however be significantly recast to represent standard data that has been implemented in varying ways across different access methods, operating systems, hardware platforms, and the five different classes of data architecture (see Section 3, Scalability and Extensibility, below)

The DISA approach only addresses the On Line Transactions Processing (OLTP) data architecture class. The practical implications of this limitation are quite important since DoD databases in these other four classes are expanding both in size and quantity faster than the size and quantity of OLTP applications.

In addition to not addressing four out of the five classes of data architecture, and because of DISA's critical error in defining its data element, the DISA approach cannot even be successful within the limited sphere of the OLTP data architecture class. That is because the DoD has virtually mandated the use of COTS (Commercial Off The Shelf) packages. It is a fact in the database market that each COTS package has its own approach to data standardization. Therefore, to achieve data standardization within a COTS, OLTP database application environment, DoD has to either totally oppose it's own COTS software acquisition policy, or force every COTS package to be rewritten before it is procured. Neither alternative is appropriate or economically feasible.

DISA's misapplication of the 1970s FIPS data field standardization guidance to its 1990s data standardization efforts prohibits a single data element from being the metadata semantic representation of many different fields in many different files. The misinterpretation naturally followed from DISA's definition of the data element that made it essentially equivalent or essentially interchangeable with a field within a file. To DISA, field and data element are in a 1-to-1 relationship. That was never efficient nor appropriate.

DISA's definition of a data element in its data standardization procedure forces the data element to be a metadata representation of a context dependent business fact instance.

In contrast, a correctly defined a data element is a metadata representation of a context-independent business fact that is used as the semantic basis for many different contained table columns, file fields, screen cells, process variables, or report fields. For DISA to be

successful with data standardization, its data element must be at a higher level of abstraction than DoD 8320.1-M-1 draft permits.

Therefore, the current DISA draft approach to data standardization, as described in DoD 8320.1-M-1, will surely be changed in the near future because 1) it is not consistent with federal standards policy, 2) it is technically inadequate for its stated purpose, and 3) it is not scalable or extensible for the full range of DoD data standardization needs.

3.0 Analysis

3.1 Federal Standards Policy

The DoD is required to follow public law (PL 104-106 and PL 104-113 (e.g., OMB Circular A-119 now public law)). These laws, and their implementation guidance in the Department of Defense, require that federal agencies use standards developed by voluntary consensus standards development organizations at the national and international level (e.g. ANSI, ISO, and IEC), to the maximum extent feasible. When however, a federal agency finds that an existing voluntary standards does not meet its needs, then government-developed standards may be used. In these cases, the federal agency is to notify the Congress, and provide an explanation of the inadequacies of existing voluntary consensus standards.

This means in practice that the DoD (Draft) Data Standardization Procedures must give way to the ISO/IEC Standard 11179 which also addresses data standardization. Only if DISA can develop an explanation of how this nationally-adopted international standard does not meet its needs, and can persuade the Congress that the voluntary consensus developed standard is inadequate, can the DoD then undertake the development of a DoD specific standard.

It is highly unlikely that DISA could develop a justification for its current data standardization procedure that will withstand a public forum scrutiny.

3.2 Adequacy for Intended Purpose

As stated above, the single most critical conceptual error in the DISA approach to data standardization is its in appropriate definition of a data element. A data element in DISA's approach is defined at the level of granularity equivalent to a column within a table. Because of this critical and methodologically fatal error, the DISA data standardization approach cannot ever achieve the objectives and goals of DoD data standardization.

The focus of data standardization must be on a level of abstraction higher than that of a "column within a table." The ISO/IEC 11179 standard, and the related ANSI X3.285 standard that is under development, correctly define data elements at this higher level of abstraction.

For sharing data among different database applications, DBMSs, operating systems, and application platforms, a data element must be defined at this higher level of abstraction so that it can exist in a one-to-many relationship with columns within tables (or any other type of container such as file, screen, process, report, etc.).

The ANSI/ISO 11179 data element is a metadata representation of a context-independent business fact that is used as the semantic basis for many different contained table columns, file fields, screen cells, process variables, or report fields. In contrast, DISA's 8320.1-M-1's definition of a data element is the metadata representation of a context dependent business fact instance.

3.3 Scalability and Extensibility

The current draft DISA data standardization approach and procedure addresses only one class of databases within an enterprise's total data architecture. The current DISA procedure does not allow DoD to share data across the full range of data bases that currently exist and that are under development.

In government and industrial enterprises there are at least five distinct classes of data architecture:

(1) Original data capture (also known as OLTP database applications),

(2) TDSA (transaction-data staging area) database applications,

(3) Subject area databases (also known as ODS (operational data store) database applications),

(4) Wholesale and retail data warehouses, and finally,

(5) Reference Data database applications.

The DISA model addresses only the first class of data architecture: original data capture (also known as OLTP). Completely unaddressed by DISA are the other four classes of data architecture. The practical implications of this limitation are economically troublesome, since the

databases within these other four classes of data architecture are expanding both in size and quantity faster than the size and quantity of OLTP applications.

In addition, because of the DISA critical error in defining the concept of a data element, the DISA approach cannot even be successful within the limited sphere of the first class of data architecture. DISA's methodology cannot achieve DoD's data standardization goals because there are valid and compelling reasons to procure COTS packages. It is a fact in the database market that each COTS package has its own approach to data standardization. DoD either has to totally oppose its own COTS software acquisition, or force every COTS package to be rewritten before it is procured. Neither alternative is appropriate or economically feasible.

The only feasible approach is to focus data standardization at a level of abstraction higher than that of a "column within a table." The focus of data standardization must be on the concept that is the semantic content of the data element. The ISO/IEC 11179 standard, and the related ANSI X3.285 standard that is under development, correctly focus on this higher level of abstraction.

4.0 Implications

DISA's critical error in its definition of a data element can be expected to cause DoD's data standardization work load to expand continuously and to be unbounded. Moreover, the rate of expansion in the data standardization workload can be expected to increase as the quantity of applications increases, rather than decrease. Thus, for every OLTP DoD data element that is standardized in accordance with the DISA approach, many more candidates for standardization will appear from the four other data architecture classes.

The approach specified in the DISA Data Standardization document is re-active rather than active. The DISA DDDS software appears to support a metadata database that is too shallow to handle the requirements of an active data standardization effort. Under the DISA reactive approach candidate data models are submitted for standardization. Given DISA's enlarging workload, standardization time will become longer and longer, notwithstanding that fact the result will be erroneous anyway. By the time the DISA data standardization process is finished, the data element specifications will be out-of-date with respect to both the data warehouse and TDSA data architecture classes of database applications. Submitting organizations would then be left with two equally unacceptable alternatives: don't submit data for standardization, or never change their original designs.

An active approach would have to both be valid and also be part of the database designer's tool kit. A valid approach is well known. As a component of the database designer's tool kit, a proposed set of standard metadata for a standardized data element would automatically be correct. The data element would be correctly specified because it would be composed from

already existing acceptable components. Once a data element is composed, it's standard metadata automatically would be available for use in the creation of columns within tables (the DISA data element) or any other type of contained fact (i.e., file fields, screen cells, process variables, or report fields). In short, under the active approach, DISA delays would be re-engineered out of the process.

5.0 Specific Comments on (draft) DoD 8320.1-M-1, Data Standardization Procedures (dated) 19 November, 1996

***Note to the Reader**: The specific comments listed below were based on the 19 November, 1996 version of "the 8320". A revised version of that document, dated 28 February, 1998 has been reviewed. The changes contained in the 28 February, 1998 version of "the 8320"document do not in anyway reduce the severity of the findings contained within this analysis. The severe problems remain.*

There are a number of errors, inconsistencies, and anomalies present in the DoD Data Standardization Procedures (DoD 8320.1-M-1) manual. These must be resolved. These problems are identified and their implications underscored in the comments below. The November 1996 draft procedures document appears to have the same fundamental errors as the previous versions of DISA data element standardization projects. The procedures and approach proposed in the document will:

1) NOT enable DoD to achieve data standardization, and

2) If followed will cause the United States Army, Navy, Air Force, and other DoD agencies to expend significant sums of monies that will ultimately be deemed wasted.

Below are just a few reasons why DoD 8320.1-M-1 is not a suitable procedure for data standardization. Please note that each and every item, individually and certainly collectively is fatal to a quality data standardization program. The items are:

1. The document only deals with standardizing OLTP databases. Given the requirement for data standardization prior to implementation, the entire concept of data warehousing is completely trashed. Data warehouses have many design changes. If each change is required to be "standardized" before implemented, the backlog of standardization will soon exceed the time to implement.

   Additionally, the concept of "alias" cross-referenced to standard reference data is missed completely. Alias was an essential part of a different data element standardization

product developed for the joint staff in the late 1980's but was not adopted by DISA. DISA's definition of data element precludes use of the alias concept.

2. The DISA document still defines data element within the context of a column of a table. This causes an infinitely expanding workload. Additionally, not all data is stored in a table. Derived data in many cases are not stored values. For example: in time-distance-rate calculators for flight data and transportation models, the results are not stored values.

3. The document only deals with columns (data elements) within tables. There are many other types of "containers" that need to have their fact representations standardized. For example, files, screens, reports, programs, etc. Given that the basic premise of DoD 8320-1-M-1 is to standardize data prior to new application implementation, then files, screens, reports, and programs also have to be subjected to the same data standardization process and rigor as is required for DISA's data elements. The costs of that effort, and the delays imposed on application development are likely to approach the DoD budget and also tens-of-thousands of staff years, respectively.

4. With respect to "domains," the concept within the document is both flawed as to its content and also is missing key components.

5. The document doesn't deal with database objects. These contain behavior, and many types of rules. Behavior must be dealt with. Object-relational DBMSs are HERE (e.g., Oracle8). Missing besides behavior are database object life cycles and state transforms.

6. The document states that candidates for standardization must be standard across functional areas. That is impossible as different functional areas VALIDLY use the same terms with different meanings. Two examples come to mind: the concept of a "track: has multiple, and equally valid meanings within DoD. A second example of are application specific 'flag codes'. These codes have no valid use apart form the application. To attempt to standardize flag codes across the entire DoD would be a gross waste of resources.

7. The document allows for waiving the standardization process. Waving MUST occur in virtually all cases because the document's strictures cannot be successfully followed. The very fact that the process can be waived is an admission that it is both fundamentally flawed and doesn't achieve the DoD data standardization goals.

   Under the DISA approach, existing on-board aircraft computer systems were designed without data standardization in mind, but still perform the functions they were designed to accomplish. The data output by these systems are equivalent to legacy OLTP COTS applications. If the DISA approach is taken, these data will forever be unstandardized. If

the business data element model is coupled with the data architecture class, TDSA, then TDSA databases can be build that contain standardized data, which, in turn can be used to build ODS databases. This approach is common within business and industry.

8. The standard in and of itself will cause significant delays because it is REACTIVE, not ACTIVE. The standard does NOT promote standardization, it only attempts to catch and "punish" violators.

   The Air Force metadata repository must be World Wide Web enabled and contain Air Force standardized metadata. A good starting point would be the Joint Command and Control meta data that was developed as part of the JOPES effort and the administrative meta data standardized for the Joint staff. The metadata repository must both available to all developers, on their critical path, and be sufficiently active to prevent non-standard metadata during the very process of creating new Air Force applications.

9. The DDDS definition of data element (i.e., a column within a table) doesn't really fulfill the need. Its meta attributes are too few (e.g., modifier classes and class word classes are woefully under specified). DISA's concept of a generic element is just another technique for specifying value set domains.

10. The meta attributes of the DISA DDDS data element have problems. E.g., there is one access name for a DBMS. Only one allowed name means that either the minimum or maximum name must be chosen. That means that in some cases the DBMSs capabilities are underused and in other cases, the DBMS itself is disallowed from use. Simply stated, the DISA concept disallows aliases.

11. The DISA DDDS data element does not contain the *common business name* meta attribute within the data element's meta attributes. Because this is missing, common business named data elements must be either treated as exceptions, or renamed to phrases that are commonly foreign to everyday users. Two examples come to mind: Telephone number, and Social Security Number.

    In the first case, Telephone Number is not a simple number. You don't add it, take its average, or perform other mathematical operations as you typically do with "numbers." Rather, Telephone Number is a compound set of subordinate numbers, each of which actually is a code. Secondly, its last word, number should actually be identifier, but according to DISA's rules that would be impossible as it is really a code. Finally, Telephone Number is not really the serial number of the telephone instrument. It is rather a different concept entirely. For some telephone-numbers, there are no telephones.

In the second case, Social Security Number too is not a simple number. It is a multi-part composite code. Additionally, Social Security Number is commonly used as an identifier. Thus, its name should really be Social Security Code, or Social Security Identifier. But since that implies that it applies to the concept of Social Security, the first two name strings are wrong also. Is "social" a modifier class word on Security. Is the concept then really a Security Identifier of a social class?

Simply stated, the DISA data element naming process implies a set of semantics based on the resulting word strings. And, the word strings have special meaning based on their appearance sequence. In the two examples above, and in many other examples, the DISA process of standardization cannot apply and if applied would result in a complete misunderstanding of the resulting name. In short, the DISA concept REQUIRES arbitrarily applied exceptions to be used.

12. Many of the meta attribute values within the DISA DDDS data element can only be known upon database application system implementation. Thus, if data standardization is required prior to implementation, then the system can never be implemented. Secondly, when the very same database application is implemented on different operating systems, DBMSs, access methods, and hardware platforms, some of the meta attribute values of the DISA data element will take on different values. Does that mean that DoD must first standardize on one DBMS, one operating system across all platforms, and on one platform for all DoD applications? Clearly, this is an impossibility.

13. The DDDS data element meta attribute, derivation code, puts composite data element's parts within "comments". What it should do is support full representation and storage of all the subparts in a bill-of-materials type application. Without a full bill-of-materials type application as the underlying processing mechanism, the quantity of DISA data elements will expand logarithmically. Like it or not, legacy systems are loaded with compound data elements, derived data, ill-defined codes and non-information bearing identifiers. For example, in one USAF project, there were more than 19,000 "DISA data elements" from more than 60 Joint Staff applications that underwent standardization. The 19,000 "DISA data elements" were standardized down to about 1200 "business data elements." Mappings between the "business data elements" to the "DISA data elements" enabled the metadata respository users to know what the data really meant regardless of where it was, how it was defined, or what it was called. Some notable mapping examples ware created for Social Security Account Number Identifer and Security Classification Code. In each example, there were hundreds of "DISA data elements" mapped to these two "business data elements."

This comment generally also applies to derived data. Finally, what about composite data that contains derived data sub-parts? Are these real-life examples just to be discarded from existence?

14. The DDDS generic element, that is really a value set domain, doesn't even have the allowed set of operations that can be specified on the identified value set.

15. The DDDS data element class word meta attribute for either data element or generic element only allows only one class word. DISA's own Figure E-2 provides a good example of why this restriction is unacceptable. For example, there clearly is a need to express an average acceleration, or minimum latitude, etc.

16. Finally, how does the data standardization analyst deal with a data element that is a number that contains embedded codes that acts as an identifier? According to DoD 8320.1-M-1 the analyst must not allow that type of reality to cloud their thinking.