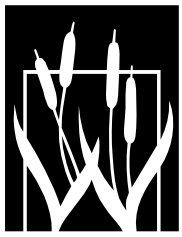

Whitemarsh
Information Systems Corporation

Data Management Program: Reference Model

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

Acknowledgments

This material is an evolution of documents that were updated during the time frame: September 2003 through December 2004. The primary contributors were Bruce Haberkamp, James Blalock, and Michael Gorman of the Office of the CIO, United States Army. The foundational components of this work has been favorably reviewed by subject matter experts within the U.S. Department of Defense.

This paper briefly explains the data management reference model that is shown in Figure 1. This model consists of the following:

- Three layers (Knowledge online, collaborative interfaces, and net centric applications)
- Three bands (Info-structure migration, data transport, and reference data)
- Four data standards (Authoritative data sources (ADS), Information exchange standards specifications (IESS), Enterprise Identifiers (EID), and XML.
- Three dimensionality arrows (Data to knowledge, data interoperability, and data community)

The top layer, knowledge online, of this data management program reference model has as its

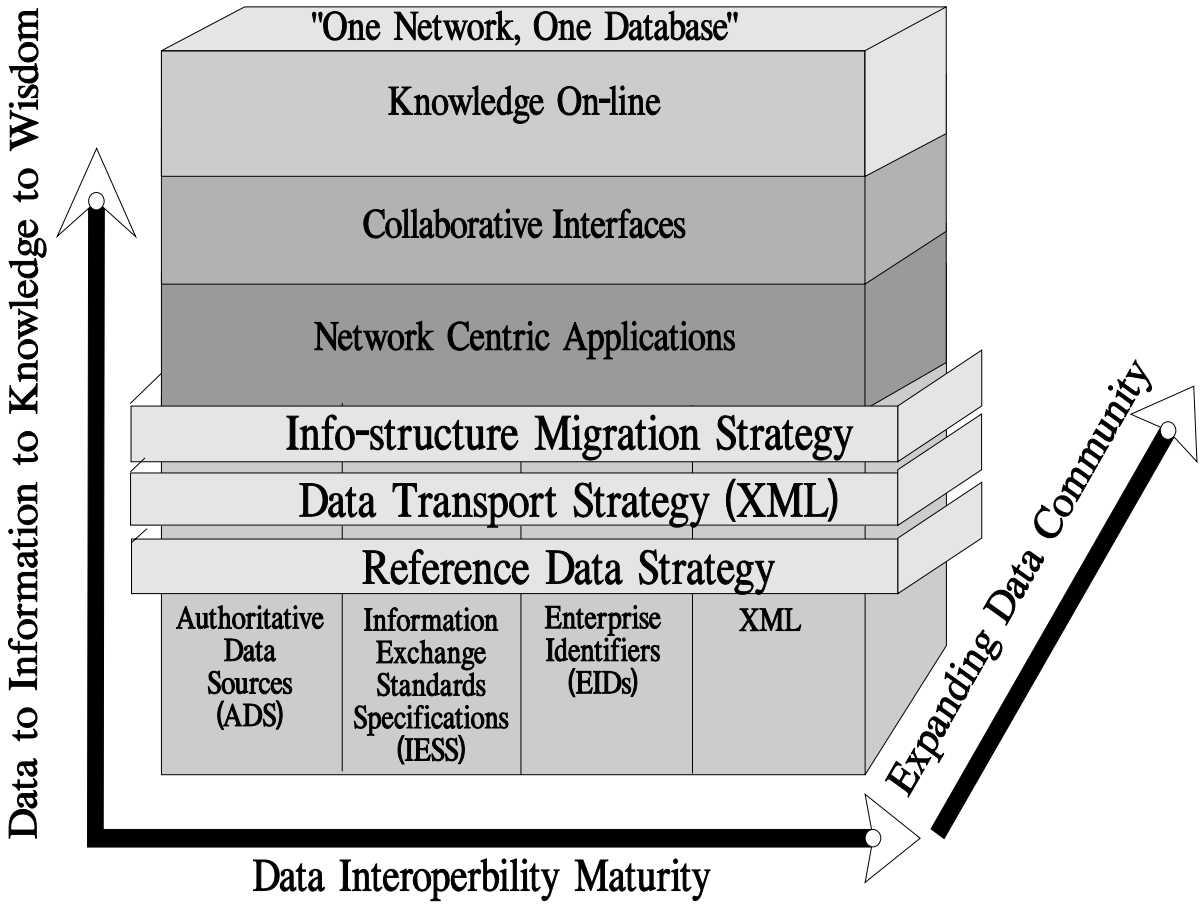

Figure 1. Data management program reference model

goal the creation of one *virtual* portal of knowledge that represents consistent, non-redundant, and semantically homogeneous data across one network and one database.. This is possible only if there is a high degree of integrated semantics across all the data that is captured, stored, and presented to users.

Integrated semantics can be achieved either through consensus based standardization and implementation through those standards, or through consensus based mapping from one set of data and semantics to another. The latter should be a transition step to the former.

Mapping is greatly eased if there are higher levels of consensus based standards. For example, suppose one legacy system has person gender with the values of M and F, and another legacy system has Sex with the values 1 and 2, if there were a higher level of abstraction with Person Gender with values Male and Female, and further, each legacy system was mapped to these values, then mapping between the two legacy systems could almost be made automatic.

Supporting the top layer is the collaborative interfaces layer that represents all the interfaces with various other systems environments. While it would be greatly eased if these systems were also based on integrated semantics, it is only required that these external semantics be mapped to existing semantics.

The next supporting layer, the environment of all net-centric applications, represents all the application information systems that are employed to capture, store, and make available the information required by the user. Ideally these applications all conform to Net Centric guidelines. That is, to the maximum extent possible they can read and write data via XML according to published XML Schemas and supporting data specifications, and these applications, to the maximum extent possible, form the basis of service oriented architectures that separate data and process from traditional bindings.

Here, it is not enough that all systems merely conform their read/write transactions into XML schemas, wrap their interchange transaction data into XML streams, and then post uniform resource locators (URLs) to where the data asset data transactions reside. If that tactic is taken, then, while connectivity (knowing and obtaining data from a data asset) can be achieved, understandability will be time consuming because the receiving system's environment staff will first have to find the appropriate XML schemas from what could be many millions, understand the XML schema, map the posting and receiving system's semantic understanding, and then construct semantic, value domain, and precision transformations prior to any meaningful data access. As stated at the outset, connectivity is not the measure of success. Understandability with minimum complexity and latency is. Employing XML without the prerequisite steps of smart and intelligent data standardization will only lead to Net Centric failure.

The three bands of the reference model, info-structure migration, data transport, and reference data represent three broad policy areas within the enterprise for standardizing the manner in

which legacy systems and data are migrated from an As-Is architecture to a To-Be architecture, and for defining and representing data to the maximum extent possible through XML specifications and standards, and the overarching approach within the enterprise for identifying and managing its critical reference data that is employed by so many different database applications.

The “data oriented” policies were derived from industry and government best practices. While these policies are stated in their “end-state” form, they can only be achieved incrementally over time. A key component of these policies is data performance planning. Simply put, data performance planning requires no more than planning for high-quality, well-engineered, best industry practice data management products and procedures. There are no “bleeding-edge” products or practices required or needed

Supporting these three policy bands are the policy specifics. These ultimately all drive the creation of net-centric data asset products.

Without these data asset products, which are intrinsically already part of other IT project failure is almost virtually assured, and significant quantities of time and money will have to be expended to resolve the lack of integration, the lack of consistent semantics, value domains, and business rules. If these products are accomplished in the manner prescribed, and if these products all reside in an integrated, enterprise-wide manner in a non-redundant federated repository environment then not only will programs have higher internal consistency and quality, they will also be easier to integrate into families of products. Further, the generation of XML related products will be quicker and more readily able to be used, thus, leading to greater understanding characterized by minimum complexity and latency.

The four data standards that form the basis for understanding-based data interoperability are” Authoritative Data Sources (ADS), Information Exchange Standards Specifications (IESS), Enterprise Identifiers (EID), and the Extensible Mark-up Language (XML). Together these four data standards are necessary to have maximum interoperability. Properly engineered and implemented, these four data standards lead to an understanding interoperability that is characterized by minimum complexity and latency.

These four data standards are commonly implemented through information systems that employ SQL DBMSs and that exchange data through formatted messages such as XML data streams. The unitary facts that are within SQL databases and exchanged through formatted messages are first defined through analysis and design, and are persistently recorded through the use of the ISO 11179 standard for data element metadata. Use of this ISO standard enables fact specification reuse many times throughout the SQL databases.

The reference model also has three dimensionality arrows. The first, “Data Information Knowledge Wisdom,” represents the transformation from data to information to knowledge to

wisdom as the four data standards (ADS, IESS, EIDs, and XML) are more and more pervasively employed throughout the enteprise as a way to have common semantics and integration with minimum redundancy.

The second dimensionality arrow, Data Interoperability, represents the increased quality and ease of creating and maintaining understanding-based data interoperability as the four components are employed across an increased quantity of database applications. Authoritative data sources leads to "one version of truth." IESSs lead to common data structures used for data exchanges that are based on and contain authoritative data sources. Enterprise Identifiers enable unique, non-redundant, and common access to all the enterprise's critical assets (real or abstract) regardless of their captive databases and systems. Access is eased through IESSs data structures based on authoritative and definitive data value sets. Finally, understanding-based data interoperability is eased if the data that is processed through the network is as non-proprietary as possible, and that now seems to be possible in an XML format.

The third dimensionality arrow, Data Community, means that these data standards, ADS, IESS, EIDs and XML, should be implemented across the entire community, which means all projects within a program, all programs within a community of interest. They should also be harmonized across all communities of interest, both horizontally and vertically. Additionally, it means that understanding-based data interoperability should be accomplished tactically, across the communities of interest, collaboratively, and across the enterprise..

The concept behind the data management program reference model is that just as a house has an architecture and requires many different components to make it complete (e.g., plumbing, wiring, carpentry, etc.), that achievement of understanding-based data interoperability also requires an architecture.