



*U.S. Federal Data Architecture Subcommittee (DAS)  
“Build to Share”*

# **Federal DAS Data Quality Framework**

**Applying Proven Quality Principles to Data Sources**

Version 1.0

October 1, 2008

# Table of Contents

<b>EXECUTIVE SUMMARY</b> .....	1
<b>SECTION 1. INTRODUCTION</b> .....	2
<b>SECTION 2. OVERVIEW OF DATA QUALITY</b> .....	5
2.1 The Business Case for Federal Data Quality.....	5
<b>SECTION 3. DATA QUALITY PERSPECTIVES IN THE FEA REFERENCE MODELS</b> .....	7
3.1 Data Quality in the PRM.....	8
3.1.1 Performance measures data validation.....	8
3.1.2 Data quality certification and benchmarks for progress.....	8
3.1.3 Information value cost chain.....	9
3.2 Data Quality in the BRM.....	9
3.2.1 Executive management accountability, data governance, data stewardship ...	10
3.2.2 Process improvements.....	11
3.2.3 Connects data creators with customers.....	11
3.3 Data Quality in the SRM.....	11
3.3.1 Focus data reconciliation at the source.....	12
3.3.2 Implement DQ as a service within transactional processes.....	12
3.3.3 Scientific methods.....	12
3.4 Data Quality in the DRM.....	12
3.4.1 Minimize the data collection burden.....	13
3.4.2 Establish enterprise data standards.....	13
3.4.3 Enterprise metadata repository.....	13
3.4.4 Designate authoritative data sources.....	14
3.5 Data Quality in the TRM.....	15
3.5.1 Improve the SDM (system development methodology).....	15
3.5.2 Optimize database performance.....	15
3.5.3 Align information architecture with data collection strategies.....	16
3.6 Conclusion.....	16
<b>SECTION 4. IMPLEMENTING THE DATA QUALITY IMPROVEMENT (DQI) INITIATIVE</b> .....	17
4.1 Developing the DQI Business Plan.....	17
4.1.1 Strategic alignment perspective.....	17
4.1.2 Alternative approaches.....	17
4.1.3 Performance improvement perspective.....	18
4.1.4 Project management perspective.....	18
4.1.5 Financial perspective.....	18
4.1.6 Information perspective.....	18
4.1.7 Change management approach.....	18
4.1.8 Next steps.....	19
4.2 Identify Data Quality Scope.....	20
4.3 Conduct Root Cause Analysis.....	22
4.4 Perform Information Value Cost Chain (VCC) Analysis.....	23
4.5 Set Data Quality Metrics and Standards.....	24
4.5.1 Key data quality measurements.....	25

---

4.6	Assess Data Against Data Quality Metrics .....	27
4.7	Assess Information Architecture and Data Definition Quality .....	28
4.7.1	Information architecture assessment .....	28
4.7.2	Data definition quality assessment.....	29
4.8	Evaluate Costs of Non-Quality Information .....	29
4.9	Develop Data Quality Governance, Data Stewardship Roles .....	30
4.10	Assess Presence of Statistical Process Control (SPC) .....	30
4.11	Implement Improvements and Data Corrections .....	31
4.12	Develop Plan for Continued Data Quality Assurance .....	33
4.13	Educate the Government Culture.....	33
4.14	Save Data Quality Products to Enterprise Metadata Repository .....	34
SECTION 5.	DATA QUALITY TOOLS.....	37
5.1	Data Profiling (Business Rule Discovery) Tools.....	37
5.2	Data Defect Prevention Tools .....	37
5.3	Metadata Management & Quality Tools.....	38
5.4	Data Reengineering and Correction Tools .....	38
APPENDIX A.	EXAMPLES OF DQI AT FEDERAL AGENCIES.....	40
A.1	Department of Housing and Urban Development .....	40
A.2	Defense Logistics Agency.....	41
APPENDIX B.	EVOLUTION OF INFORMATION QUALITY MANAGEMENT .....	43
APPENDIX C.	GLOSSARY .....	44
APPENDIX D.	ADDITIONAL REFERENCES .....	48

---

## EXECUTIVE SUMMARY

Data quality improvement initiatives provide a framework for federal agencies to:

- Target the spending of scarce data quality resources by identifying data used across organizational boundaries to meet high-profile business performance reporting responsibilities,
- Document key data validation, extraction, and transformation processes to ensure repeatability and efficiency in the data management of mission-critical data,
- Implement data quality standards for systems and data supporting high-profile business performance reporting responsibilities, and
- Implement a methodology for independent verification of high priority, performance-measurement information.

Accurately reporting an agency's performance goals and objectives may require the development of new data systems and the fixing of old ones. A data quality improvement program can assist agencies to make informed choices between the "old" (legacy) and the "new", by identifying where the most definitive and precise performance information on the accomplishment of agency-wide program goals exists.

Obtaining senior management support by means of a detailed data quality business plan is essential to sell the data quality value proposition to federal agencies and other communities of interest. Federal data quality projects will gain traction if executives institute incentive programs to encourage employees to follow the new data quality policies, and if the agencies publicly recognize employees who make major contributions toward the data quality improvement process.

To ensure high quality of data within federal agencies' information systems, data quality activities must provide agencies with repeatable processes for detecting faulty data, establishing data quality benchmarks, certifying (statistically measuring) their quality, and continuously monitoring their quality compliance. The ultimate outcome of ongoing data quality monitoring efforts is the ability to reach and maintain a state in which government agencies can certify the quality level of their data. This will assure the government agencies' data internal and external consumers of the credibility of information upon which they base their decisions.

A very deep appreciation goes to the DAS working group responsible for the development of this document:

Mark Amspoker, Citizant, Inc.  
Shula Markland, HUD  
Ryan Day, USDA  
David Loshin, Knowledge Integrity, Inc.  
Richard Ordowich, Knowledge Integrity, Inc.

## SECTION 1. INTRODUCTION

As federal agencies transform to become more citizen-centered and results-oriented, they are likely to face added demands for data access. At the same time, reduced resources may encourage decisions to consolidate and eliminate systems, and agencies may look to increased sharing opportunities. Data sharing and system consolidation occurs through system integration, data migration, and interoperability. As a result of data sharing and system consolidation, agencies often discover that different business uses of data impose different quality requirements, and that data that were of acceptable quality for one purpose may not be acceptable for other purposes. For example, data that were of sufficient accuracy and timeliness for local use may not be acceptable when used in a broader community. Costs of inaccurate or inadequate data can be steep, resulting in tangible and intangible damage ranging from loss of information consumer confidence to loss of life and mission.

Data quality management in the federal government is focused on the same problems and issues that afflict the creation, management, and use of data in other organizations. The lack of data integration due to incompatible database structures, poor quality and integrity of data, and inconsistent data standards hinders the collection, manipulation, and transmission of information within a community of interest.

Managing data quality is essential to mission success. It ensures that:

- Data are managed as a national asset,
- Data support effective decision-making, and
- The right data reach the right person at the right time in the right way.

Improving data quality will lower automated support costs by streamlining information exchange and increasing information sharing reliability.

In this Federal Data Architecture Subcommittee (DAS) Data Quality Framework (“DAS DQ Framework”), data quality is described as a series of disciplines and procedures to ensure that data are meeting the quality characteristics required for use in communities of interest (COI). The DAS DQ Framework defines approaches for people, processes and technology that are based on proven methods, industry standards, and past achievements.

This document can be viewed within the context of the objectives laid out in the Office of Management and Budget’s (OMB) final government-wide Information Quality Guidelines (OMB 67 FR 8452). Those Guidelines implemented Section 515 of the Treasury and General Government Appropriations Act of Fiscal Year 2001 (Public Law 106-554; H.R. 5658) (“Section 515”), which directed OMB to issue guidelines that “provide policy and procedural guidance to federal agencies for ensuring and maximizing the quality, objectivity, utility, and integrity of information (including statistical information) disseminated by Federal agencies.” The Government-wide Information Quality Guidelines<sup>1</sup> issued by OMB in response to Section 515 define information as “any communication or representation of knowledge such as facts or data, in any medium or form, including textual, numerical, graphic, cartographic, narrative, or audiovisual forms”<sup>2,3</sup>. The Government-wide Information Quality Guidelines (IQ Guidelines) define “dissemination” as agency initiated or sponsored distribution of information to the public.

---

<sup>1</sup> Add reference to Feb 22, 2002 FR notice

<sup>2</sup> This definition includes information that an agency disseminates from a web page, but does not include the provision of hyperlinks to information that others disseminate.

<sup>3</sup> This definition does not include opinions, where the agency’s presentation makes it clear that what is being offered is someone’s opinion rather than fact or the agency’s views.

Of particular relevance to the DAS DQ Framework, these IQ Guidelines state that:

- Overall, agencies shall adopt a basic standard of quality (including objectivity, utility, and integrity) as a performance goal and should take appropriate steps to incorporate information quality criteria into agency information dissemination practices. Quality is to be ensured and established at levels appropriate to the nature and timeliness of the information to be disseminated. Agencies shall adopt specific standards of quality that are appropriate for the various categories of information they disseminate.
- As a matter of good and effective agency information resources management, agencies shall develop a process for reviewing the quality (including the objectivity, utility, and integrity) of information before it is disseminated. Agencies shall treat information quality as integral to every step of an agency's development of information, including creation, collection, maintenance, and dissemination. This process shall enable the agency to substantiate the quality of the information it has disseminated through documentation or other means appropriate to the information.

The IQ Guidelines required that each agency develop its own, agency-specific guidelines. In many agencies, these guidelines have served to highlight the importance of quality of the underlying data bases. As further progress is made in implementing the agency-specific IQ Guidelines, additional improvements are expected in data quality.

The DAS DQ Framework applies specifically to the creation, collection, and maintenance of data used in an agency's information-development process; that is, it refers to the business processes surrounding the use of federal data stored in internal authoritative data sources (ADS), some of which may not be available to the public. An agency's quality policies and procedures should be designed to ensure that internal data and data systems are of appropriate quality for its intended use, taking into account the possibility that information derived from those data may eventually be disseminated. OMB's definition of "disseminated" encompasses information which has the appearance of representing agency views. This includes internal or third-party information that is used in support of an official position of the government entity, as well as publicly available analyses of internal data. Thus, the quality of ADS may sometimes have important implications for the information upon which public policy is based.

This document embraces the principles upon which the IQ Guidelines are based. Both the DAS DQ Framework and the IQ Guidelines embrace the development of processes for reviewing data quality, and both recognize that high quality comes at a cost and agencies should weigh the costs and benefits of higher information quality. The principle of balancing the investment in quality commensurate with the use to which it will be put is generally applicable to all data that the federal government generates.

OMB defines "quality" in terms of utility, objectivity, and integrity. The DAS DQ Framework provides granularity to the meaning of "data quality" when specifically applied to ADS. This document introduces terms that characterize important quality dimensions of ADS data, including timeliness, accuracy, completeness, consistency (data content quality dimensions), accessibility, contextual clarity, and usability (data presentation quality dimensions). Table ES1 below maps these terms to the terms used in the Information Quality Guidelines.

<b>OMB Information Quality Dimensions</b>	<b>OMB Definition from final government-wide IQ Guidelines</b>  ( <a href="http://www.whitehouse.gov/omb/fereg/reproducible2.pdf">http://www.whitehouse.gov/omb/fereg/reproducible2.pdf</a> )	<b>DAS DQ Framework Granular Measures Supporting OMB Guidelines (for definitions see Section 4.5.1)</b>
Utility	Utility refers to the usefulness of the information to its intended users, including the public.	Timeliness, Concurrency, Precision, Accessibility,

		Contextual Clarity, Rightness and Usability.
Objectivity	<p>Objectivity involves two distinct elements, presentation and substance.</p> <p>The first involves whether disseminated information is being presented in an accurate, clear, complete, and unbiased manner. Here the focus is on the context in which the data are presented as well as the associated documentation.</p> <p>The second focuses on the accuracy, reliability, and potential for bias in the underlying information, including whether the original data and subsequent analysis were generated using sound research and/or statistical methods.</p>	Accuracy to Reality, Accuracy to Surrogate Source, Precision, Validity, Completeness, Relationship Validity, Non-duplication, Consistency, Concurrency, Contextual Clarity, Usability and Derivation Integrity.
Integrity	The Guidelines use a definition of integrity that refers specifically to the security of information. In this instance, integrity refers to the protection of the information from unauthorized access or revision, to ensure that the information is not compromised.	Data security is not assessed in the processes of the DAS DQ Framework.

**Table ES1 - OMB IQ Dimensions Mapped to Granular Dimensions in DAS DQ Framework**

The impact of data quality initiatives can go beyond data management and information exchange improvements. They can provide direct support in the development of Federal Enterprise Architecture (FEA) reference models. Like data quality improvement, Enterprise Architecture development establishes a clearer line of sight from investments to measurable performance improvements whether for the entire enterprise or a segment of the enterprise. In Section 3 of the DAS DQ Framework, core data quality principles are displayed alongside the FEA reference models where appropriate to buttress the case for implementing a data quality improvement program at the federal level. This guidance assists architects to develop and use segment architecture to:

- Describe the current and future state of the agency and its segments,
- Define the desired results for each segment,
- Determine the resources needed for an agency's core mission areas and common or shared services,
- Leverage resources across the agency, and
- Develop a transition strategy to achieve the desired results.

The DAS DQ Framework provides the means for embedding industry-proven data quality procedures and practices into agency business processes.

The structured Data Quality Improvement (DQI) initiative articulated in Section 4 of this document can reap substantial benefits to federal agencies and COI's that wish to embark on a data quality program or make improvements in their existing quality systems. The activation of such a program, appropriately tailored to an agency's size and budget, deserves to be effectively communicated to business managers who will sponsor both the technology and the organizational infrastructure in order to ensure a successful program. By no means is the DQI methodology introduced in this document the only possible set of procedures to bring about significant improvement in federal data quality. However, the thirteen DQI process steps outlined in Section 4 represent best practices that have been implemented at a number of federal agencies with great success (see Appendix A for two examples of successful federal DQI).

## SECTION 2. OVERVIEW OF DATA QUALITY

**"The degree to which the data/information is fit for use for the task at hand in terms of dimensions such as timeliness, completeness, and believability." (Dr. Richard Wang)**

The definition of data quality has evolved over the past half century. Prior to the 1970's, data quality usually referred to "the degree of excellence of data." Data were of excellent quality if they were stored according to data type, if they were consistent and not redundant, and if they conformed to prescribed business rules. During the 1990's, however, a number of data quality thought leaders began to take the quality principles of Dr. W. E. Deming, W. Shewhart, P. B. Crosby and M. Imai (for a brief discussion of the evolution of information quality management refer to Appendix B) and adapt them to information management with the same results. Information is a product "manufactured" by one or multiple processes (taking a loan or a grant application) and consumed by other processes (reporting performance indicators) or customers (public housing authorities).

Today, J.M. Juran's definition of data quality is thought to be definitive: "Data are of high quality if they are fit for their intended uses in operations, decision making and planning." Larry English writes that "Information (i.e., data in context) quality means consistently meeting the information customer's expectations." Thomas Redman, another data quality thought leader, says that "Data are of high quality when data are relevant to their intended uses, and are of sufficient detail and quantity, with a high degree of accuracy and completeness, consistent with other sources, and presented in appropriate ways."

The terms data and information are often used loosely as though they are interchangeable. In the IQ Guidelines, data and facts are included in the broader definition of 'information.' The DAS DQ Framework focuses only on the subset of information referred to as data. In this document, data are defined as single representations (units) of fact that may later be used as the raw material in a predefined process that ultimately produces a higher level information product. This document does not directly address the meaning given to data or the interpretation of data based on its context, although those later uses should dictate the level of quality of the data themselves.

Data quality does not happen by accident. Agencies and COI's must establish standards and guidelines for all personnel to follow to ensure that data quality is addressed during the entire lifecycle of data's movement through information systems. Data quality cannot long endure without the establishment of standards for defining the data, naming the data, developing domain (valid values) and business rules, and modeling the data. Data quality should include guidelines for data entry, edit checking, validating and auditing of data, correcting data errors, and removing the root causes of data contamination. Standards and guidelines should also include policies and procedures, such as operating procedures, change-control procedures, issue management procedures, data dispute resolution procedures, roles and responsibilities, and standard documentation formats. All of these policies, procedures and definitions are part of the framework for data quality.

### **2.1 The Business Case for Federal Data Quality**

When Congress passed the Government Performance and Results Act of 1993 (GPRA), it signaled to the nation that it wanted the federal government to change the way it was doing business. Instead of measuring the success of departments and agencies solely by looking at how well they implement their programs, Congress wanted to know the results, or outcomes, that accrued from departments' and agencies' efforts.

GPRA challenged government managers to define their agency's impact on the lives of the American people. These expected impacts were to be stated as long-term, outcome-oriented goals in a five-year



strategic plan. Once the long-term outcome goals were defined, agencies were to develop annual outcome-oriented goals that would stand as the building blocks for meeting the long-term goals.

In general, GAO audits over the past several years have concluded that federal agencies struggle with developing a comprehensive approach to the quality of disseminated information because of internal management shortcomings, the complexity that results from the size and scope of federal agencies and departments, and the need to standardize and modernize technology and information technology (IT) processes. These audits have concluded that improved data quality calls for a more organized and sustained approach at federal agencies, requiring a long-term commitment. As part of the establishment of such a program, data quality principles -- including data definition standards, data security and privacy guidelines, data modeling guidelines, and data management infrastructure and policies -- should also be resolved in alignment with emerging enterprise architecture practices.

To ensure high quality of data within federal agencies' information systems, a data quality process must provide agencies with a systematic, industry-proven, repeatable process for detecting faulty data, establishing data quality benchmarks, certifying (statistically measuring) their quality, and continuously monitoring their quality compliance. It is important that the concept of continuous monitoring of data quality be both understood and adhered to for a successful data quality effort. Although reaching required quality levels is a major achievement for a business environment, this should not be construed as the end of data quality efforts for that environment. Once the state of data quality is reached, it needs to be maintained. Continuous monitoring is the mechanism by which agencies can manage the quality of their data with the ever-present possibility of data corruption.

The ultimate outcome of ongoing systematic efforts is the ability to reach and maintain a state in which government agencies can certify the quality level of their data. This will assure the government agencies' data internal and external consumers of the credibility of information upon which they base their decisions.

Enterprise-wide data management must be developed, implemented, and enforced in the federal government to improve data quality in a holistic, cross-program way. Because data quality improvement is a process and not an event, the following enterprise-wide disciplines should be phased in and improved upon over time:

- A stronger personal involvement by management,
- High-level leadership for data quality,
- New performance evaluation measures based on data quality,
- Data quality enforcement policies,
- Data quality assessments,
- Additional training for data owners and data stewards about their responsibilities,
- Data standardization,
- Metadata and data inventory management techniques, and
- A common data-driven methodology.

## SECTION 3. DATA QUALITY PERSPECTIVES IN THE FEA REFERENCE MODELS

The Federal Enterprise Architecture (FEA) is an initiative of the OMB that aims to comply with the Clinger-Cohen Act and provide a common methodology for information technology (IT) acquisition in the United States Federal government. The primary purpose of the FEA is to identify opportunities to simplify processes and unify work across agencies and within similar lines of business of the federal government, leading to a more customer-focused government that maximizes technology investments to better achieve mission outcomes.

The FEA is a collection of reference models that develop a common taxonomy for describing IT resources. These include the Performance Reference Model (PRM), the Business Reference Model (BRM), the Service Component Reference Model (SRM), the Data Reference Model (DRM) and the Technical Reference Model (TRM). The five models are designed to be interrelated and mutually supporting – their purpose is to facilitate cross-agency collaboration in support of citizen-focused delivery of services.

Data quality principles and initiatives can enable better delivery of these services at each FEA level, as shown in the following graphic (Figure 3-1):

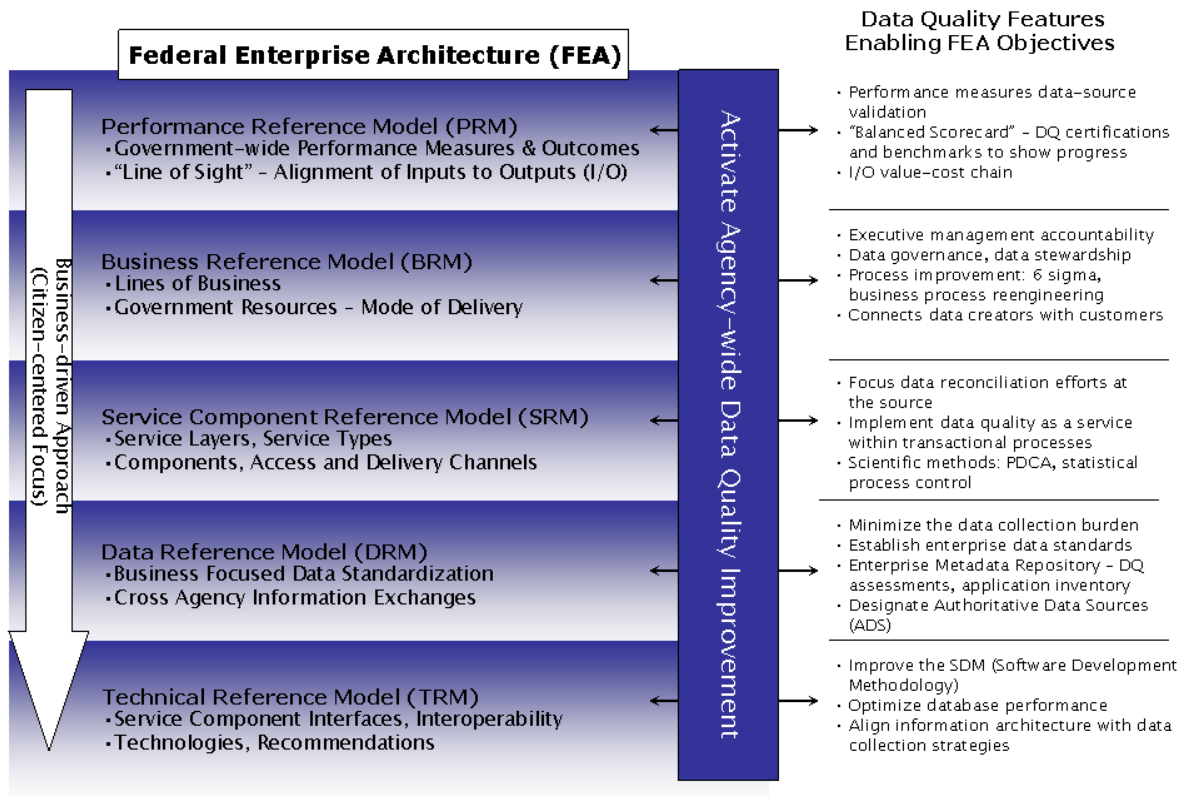
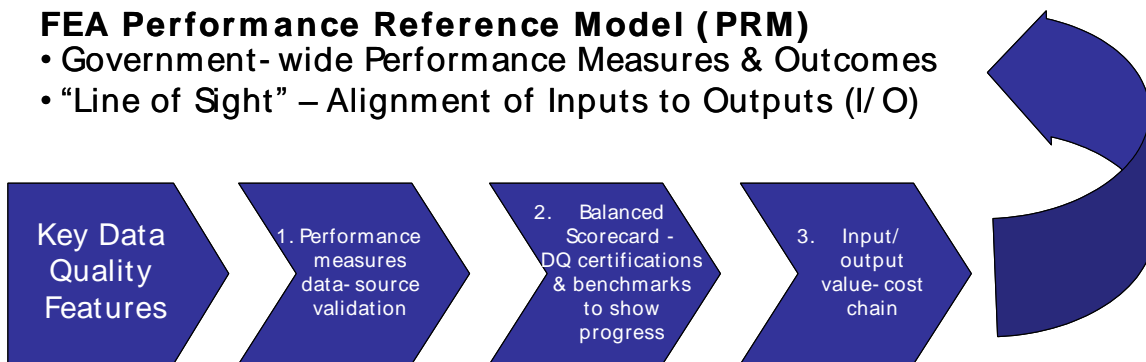


Figure 3-1: The FEA-Data Quality Value Proposition

### 3.1 Data Quality in the PRM

The PRM is a standardized framework to measure the performance of major IT investments and their contribution to program performance. By utilizing a number of existing approaches to performance measurement, the PRM identifies performance improvement opportunities that span traditional organizational structures and boundaries.



**Figure 3-2: Key-Data Quality Features Supporting PRM Compliance Guidelines**

#### 3.1.1 Performance measures data validation

Data quality initiatives enable federal agencies to meet high-profile business performance reporting responsibilities, such as the Annual Performance Plans (APP) now required of agencies through the Government Performance and Results Act. The business performance measurements specified in the APP require a detailed data source discussion. Data quality initiatives can support and validate the data discussion by:

- Assisting business areas in getting to the right data from the right information systems at the beginning of the performance reporting process,
- Documenting the data validation process to ensure repeatability and efficiency,
- Implementing data quality standards for systems and data supporting the performance measurements, and
- Implementing a methodology for independent verification of high priority, performance-measurement data.

Effectively reporting an agency’s performance goals and objectives may also require developing some new data systems and fixing old ones. A data quality program can assist agencies make informed choices between the “old” (legacy) and the “new”, by identifying where the most definitive and precise performance information on the accomplishment of agency-wide program goals exists.

#### 3.1.2 Data quality certification and benchmarks for progress

At the PRM level, data quality initiatives provide effective benchmarking of reported results (Certification) after the processes that produce or maintain high-performance data are improved and the data not meeting agreed-upon standards have been corrected. Certification:

- Assesses whether the data produced by create and maintain processes are in compliance with the definition and quality standards of the data, and

- Assesses whether the data contained in files, databases, data warehouses, data marts, reports, and screens are also in compliance.

Based on observations and findings, data quality processes recommend improvements to the procedures used to implement data quality best practices (defect prevention), as well as improvements in data correction procedures, for performance-measurement data.

### 3.1.3 Information value cost chain

Data quality's Information Value Cost Chain (VCC), also referred to as an Information Product (IP) map, lends factual evidence to support the difficult task of estimating the value and performance of government IT investments (see Section 4.4 for a continued discussion and examples). This process maps the data's complete life cycle to include the logistics of their creation, input into an original information system, the steps of their transformation into a "finished" IP, and the logistics of their output to the customer. The process also includes detailed descriptions of the servicing of the data (their maintenance as well as support to customers using the data). Costs are attached to the data at each stage of their life cycle. These costs can then be compared against the real and intrinsic value of the data to support the federal agency's "bottom line", in this case the agency's adherence to a 5-Year Strategic Plan, APP performance goals, or other key objective.

IP's that do not yield a profit (i.e., their costs of production and maintenance over their life cycle exceed their value to the agency's bottom line) would be prime targets for reprocessing. Data's "profit margin" gives federal agencies improved line of sight into the efficiency of their technology and provides important feedback to key federal supporting business areas:

- Procurement: producing the means and materials to acquire the data.
- Human resources: allocating the personnel required to support the data.
- Technology development: technologies to support value-creating data.
- Agency infrastructure: organizational structure, control systems, culture and business environment.

## 3.2 Data Quality in the BRM

The BRM provides a framework that facilitates a functional (rather than organizational) view of the federal government's lines of business, including its internal operations and its services for citizens, independent of the agencies, bureaus and offices that perform them. Data quality initiatives support the BRM framework by encouraging the data originators and data consumers who are integral to the smooth functioning of their line of business to become more involved in the business context and conditions. Data archeology (discovering data through forensics), data cleansing (correcting bad data), data quality enforcement (preventing data defects at the source), and knowledge of authoritative data sources (ADS) should be business objectives. Therefore, data quality initiatives are business initiatives and require the integration of technical people with business people. Business management should be in alignment with proven quality management practices, and data quality management activities can support these practices.

## Business Reference Model (BRM)

- Lines of Business
- Government Resources – Mode of Delivery

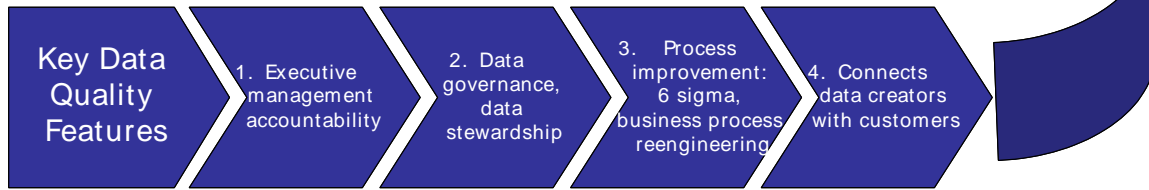


Figure 3-3: Key-Data Quality Features Supporting BRM Compliance Guidelines

### 3.2.1 Executive management accountability, data governance, data stewardship

Data quality initiatives will ultimately fail without strong executive sponsorship at the business level. Without this sponsorship from the top level, the data quality policies of the agency and the work habits of the staff will not change. Therefore, supporting the BRM through data quality initiatives means establishing data governance groups that are staffed with independent validation and verification (IV&V) data quality assessment experts, internal data administrators, metadata administrators, and data quality stewards:

- **Data Quality Group**—A (usually) independent IV&V branch that establishes and maintains a data quality handbook pertinent to the agency involved; conducts periodic DQ assessments based upon a rigorous and defined methodology; maintains the list of information systems/products targeted for DQ review; establishes and monitors timeframes for DQ reviews; develops and provides overall DQ reporting; conducts DQ training to the enterprise; tracks organizational error trends; and may also develop DQ marketing/incentive/promotional efforts. It is the responsibility of the Data Quality Group as well to audit metadata and data models, and to be involved in data reconciliation efforts by helping to identify and resolve the root causes of data quality issues. The findings of the audits and reconciliation efforts should feed back into a continuous data quality improvement cycle.
- **Data Administrators**—These individuals are responsible for establishing the linkages between source data, information products and policies or regulations that enforce how data and information can be used, and for establishing information policy. Often the senior individuals at the program office level, Data Administrators have ultimate responsibility for ensuring accuracy, completeness, validity and reproducibility of data stored in systems used to support the program office lines of business.
- **Metadata Administrators**—These individuals are responsible for loading, linking, managing, and disseminating metadata to facilitate the common understanding of data and to encourage data reuse. They are responsible for the enterprise logical data model, for establishing and maintaining naming standards, and for capturing data-related business rules. They are accountable for the quality of data in all data repositories that support the program office; maintenance of data models, database design and data definitions. Metadata Administrators are accountable for knowledge about the program office value and cost chains for information systems and data elements that are required for reporting key business processes. Duties include: (1) producing and updating Data Element Dictionaries using industry standard Computer Aided Software Engineering (CASE) tools; (2) ensuring that configuration management software, where integrated into business systems, is used to maintain version control for mission-critical data elements supporting the agency's mission; and (3) maintaining a test database of results for CM and quality assurance (QA) purposes.

### 3.2.2 Process improvements

The BRM provides an organized hierarchical construct for describing the day-to-day business operations of the federal government using a functionally driven approach. Data quality initiatives provide greater visibility into these business operations through their built-in process improvements:

- Product specification (customer input and solicitation),
- Continuous process improvement (CPI)/6 Sigma, and
- Business process re-engineering (BPR).

It is important to recognize that existing agency business processes could already have in place some form of quality checks and balances, regardless of whether or not these are termed data quality. Section 4 will discuss some of the ways formal data quality initiatives can learn about and tap into existing quality processes, ultimately strengthening their reach and effect throughout the enterprise.

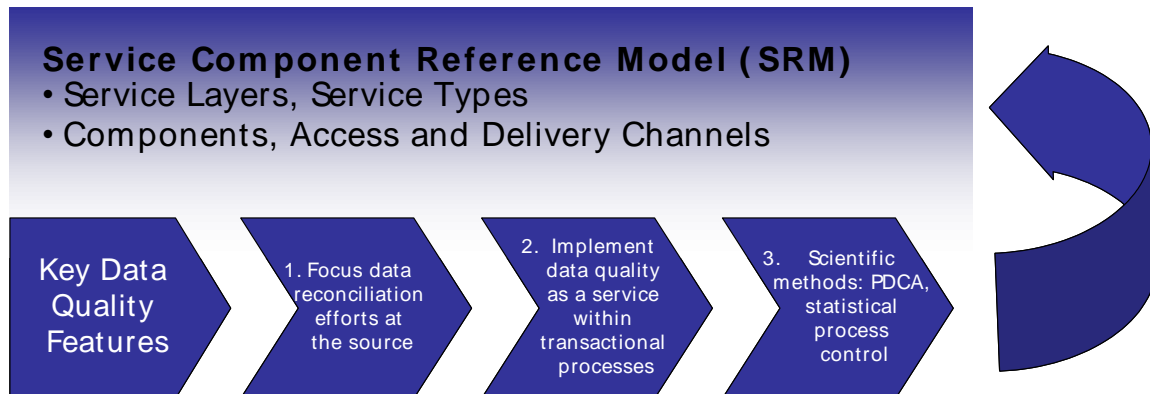
### 3.2.3 Connects data creators with customers

If the COI's and federal agencies are to provide high quality shared data, they must first understand their customers. The processes of data creation, maintenance, propagation and delivery involve multiple customers; these customers have multiple needs and expectations that the data must meet. Information customers can be internal or external. Internal customers are processes and people consuming data to make critical decisions, such as underwriting an application or securing funding for future programs, providing insight into an agency's performance, or servicing the public. External customers include the public, state and local governments, Congress, public service organizations, and the Executive Branch.

Building requirements and feedback channels between these creators and customers of data is one of the most effective ways that data quality principles can support the BRM.

## 3.3 Data Quality in the SRM

The SRM is intended to support the discovery of government-wide business and application service components in IT investments and assets. The SRM is structured across horizontal and vertical service domains that, independent of the business functions, can leverage the reuse of applications, application capabilities, components, and business services.



**Figure 3-4: Key-Data Quality Features Supporting SRM Compliance Guidelines**

### 3.3.1 Focus data reconciliation at the source

Data quality initiatives aid an agency's desire to automate customer service and business management services by focusing the creation of clean, high-quality data at the source. Automation inevitably means data reconciliation of formerly manual processes, and reconciliation through data quality best practices is easier and cheaper to perform than simply making data corrections. Reconciliation is the process of capturing, storing, extracting, merging, separating, copying, moving, changing, or deleting data. This is especially true for data warehouse applications that extract data from multiple operational source files and merge the data into one target database. If a business area has adopted an architected data mart strategy, then the various data marts also have to be reconciled to each other to guarantee consistency. This includes having one central staging area with extensive reconciliation programming for every input-process-output module.

### 3.3.2 Implement DQ as a service within transactional processes

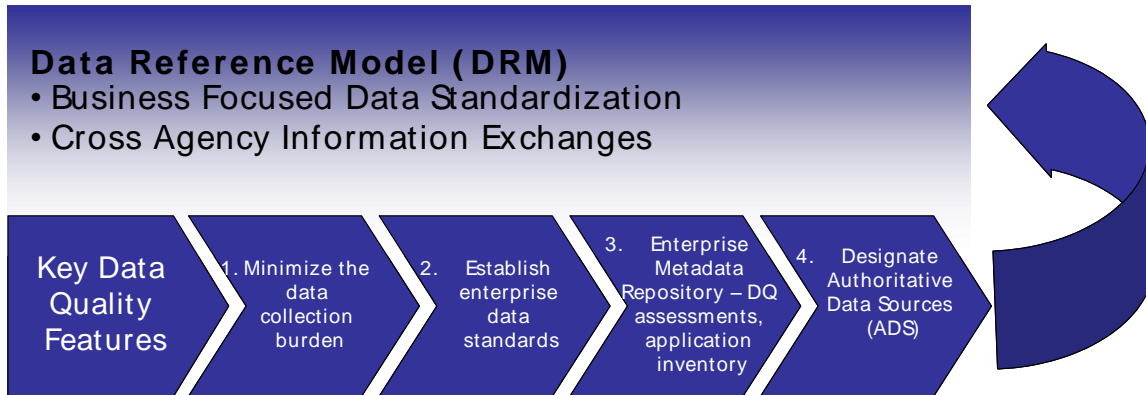
At the SRM level, data quality initiatives provide further benefit to transactional business services – in real-time – by validating and/or correcting new data, as well as automating the rationalization and standardization of data from different countries (i.e., internationalization of data). Data quality deployments at the federal level are increasingly addressing “operational” data, i.e., the data in the systems that drive day-to-day operations at the agency. The shift to operational data quality also increases the need for the data quality environment to interoperate with the overall enterprise IT environment in a seamless, service-oriented manner. As the Federal Government increasingly turns to SOA (Service-Oriented Architecture), agencies are seeking data quality solutions that can deliver data services adaptively and “on the fly”: via multiple protocols and platforms that are easily consumable by a wide variety of downstream needs.

### 3.3.3 Scientific methods

The SRM is additionally supported by data quality's built-in scientific methodologies, including statistical process control. Statistical process control uses statistical methods (i.e., run charts) to monitor the actual quality of data consumed by a business process over a defined period of time. Clarifying the initial quality requirements for data through statistical process control is the key to service at the operational level. Quality Assurance Plans are developed as part of this process control to ensure that the outputs of transactional systems are correct and credible; that the system's personnel continually improve on the knowledge and skill sets necessary to run their business; and that data are gathered throughout the system's life cycle to improve performance at every level as well as improve the allocation of key resources.

## 3.4 Data Quality in the DRM

The DRM categorizes government information into greater levels of detail. It also (1) establishes a classification for federal data; (2) streamlines information exchange processes (e.g., the use of technologies such as XML/XSD in the data sharing/information exchange package area); (3) identifies duplicative data resources; and (4) describes artifacts which can be generated from the data architectures of federal agencies. The DRM's three principal areas of standardization are Data Description, Data Context, and Data Sharing.



**Figure 3-5: Key-Data Quality Features Supporting DRM Compliance Guidelines**

### 3.4.1 Minimize the data collection burden

Data quality initiatives provide improved strategies and protocols for data collection in support of an agency’s mission. Through the analysis of VCCs and other DQI artifacts (see Section 4), protocols can be designed to minimize the data collection burden frequently placed on agency customers and business partners. Efficiencies of data collection can be obtained by requiring that data elements be gathered and documented in a manner that produces information consistent across agency lines of business. At the other end of the information product chain, developing consistent reporting protocols will maximize the utility of the data for identifying the successes of each individual line of business.

### 3.4.2 Establish enterprise data standards

A key contribution to the DRM is the steady progression toward enterprise data standards through the artifacts and general cultural effects of data quality improvement. Data quality initiatives ferret out inconsistencies in data naming standards across the enterprise and begin the journey towards the application of consistent standards, such as the convention of name compositions using prime words, qualifiers or modifiers, and class words. Data administrators are usually trained in the various industry-standard naming conventions. The goal is for agencies to publish a standard enterprise-wide list that includes industry-specific and organization-specific terms. Once established across the enterprise, data standards can better support agency data integrity, accuracy and objectivity, as well as enable data sharing between diverse COIs.

Similar to what was stated above in Section 3.2.2, existing agency business processes could already have in place a robust Data Management program, regardless of whether or not formal data quality assessment are included. The agency may have already established and published enterprise-wide data standards and applied them to agency day-to-day business operations. Existing enterprise data standards, guidelines and assets may include unique data object naming and definition standards, data model class and attribute standards, conceptual (logical and physical) data model standards, extensible markup language (XML) naming and design rules, and web services standards and guidelines. Section 4 of the DAS DQ Framework will give formal data quality approaches that can be embedded into an existing Enterprise Data Management program.

### 3.4.3 Enterprise metadata repository

Outputs of data quality initiatives become the basis for an Enterprise Metadata Repository (EMR), wherein the agency’s enterprise metadata are inventoried and stored. Metadata is “descriptive



contextual information about architectural components” and can be business metadata, technical metadata, process metadata, and usage metadata. Large amounts of business metadata can be collected about business functions, business processes, business entities, business attributes, business rules, and data quality. Technical metadata represents the physical architectural components, such as programs, scripts, databases, tables, columns, keys, and indices. Process metadata describes any type of program logic that manipulates data during data capture, data movement, or data retrieval. Usage metadata is statistical information about how systems are used by the business people: for example, what type of data is accessed, by whom, how often, and for what purpose.

The EMR should be set up in such a way that it supports the standards for metadata capture and usage. The EMR is an essential tool for standardizing data, for managing and enforcing the data standards, and for reducing the amount of rework performed by developers or users who are not aware of what already exists and therefore do not reuse any architectural components.

For the purpose of finding redundant and inconsistent data, logical entity-relationship modeling with complete data normalization is still one of the most effective techniques because it is a business analysis technique that includes identification, rationalization, and standardization of data through business metadata. Because every business activity or business function uses or manipulates data in some fashion, a logical data model documents those logical data relationships and the business rules, regardless of how the data or the functions are implemented in the physical databases and applications. UML (Unified Modeling Language) class diagramming is another effective technique for representing an enterprise information model.

Logical data models created for individual applications should be merged into one cohesive, integrated enterprise logical data model. This activity is usually performed by the data administration department, which might be part of the Data Quality Group. The enterprise logical data model is the baseline business information architecture into which physical files and databases are mapped. Agencies should establish standards for creating logical data models as part of system development activities and for merging the models into the enterprise logical data model.

#### 3.4.4 Designate authoritative data sources

**"Create once, store once and update once to then use many times."**

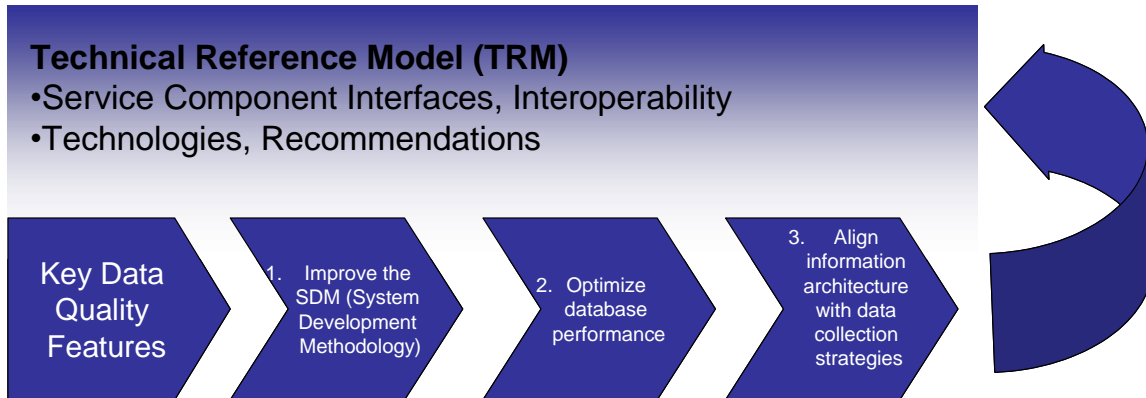
The EMR gives an agency a further advantage for determining Authoritative Data Sources (ADS). An ADS can be defined as a cohesive set of data assets that provide trusted, timely and secure information to support a business process. Identifying the best data source without regard to ADS can be time consuming and expensive: if there are multiple versions of the same data source, then the cost of cycling through all of them to determine the most correct version can put a strain on agency resources. Through better metadata management obtained via the EMR, ADS-search is automated, resulting in:

- Reduction of knowledge acquisition time,
- Identification of “best” initial data products,
- Discovery of intended purpose of data clearly and concisely, and
- Achievement of reliable and secure metadata configuration management.

While automated ADS-search can be a boon to an agency’s business mission, the development of its technology must be considered in terms of costs vs. cost savings. In addition, while there may be clear expectations that data from one source are better than another, this reassurance may ultimately be immaterial if the data cannot be assumed to be trusted and reliable. The need for “pedigreed data” (knowledge of the provenance of data) will continue to be important with regard to ADS development.

### 3.5 Data Quality in the TRM

The TRM is a component-driven, technical framework used to categorize the standards, specifications, and technologies that support and enable the delivery of service components and capabilities. It provides a foundation to categorize the standards, specifications, and technologies to support the construction, delivery, and exchange of business and application components (Service Components) that may be used and leveraged in a component-based or SOA environment.



**Figure 3-6: Key-Data Quality Features Supporting TRM Compliance Guidelines**

#### 3.5.1 Improve the SDM (system development methodology)

Data quality checks should be integrated and standardized within the agency’s SDM in support of the TRM. A development methodology is a common roadmap that provides a complete list of all the major activities and tasks to be performed on projects. The trouble with traditional methodologies is that they do not support cross-organizational data integration activities because operational systems were rarely designed with integration in mind. But increasing demand for integrated systems requires a new type of data-driven methodology that includes the appropriate data quality improvement tasks. For example, the methodology must have a separate development step for incrementally building the enterprise logical data model and enforcing data standardization across all projects.

#### 3.5.2 Optimize database performance

Data quality initiatives typically include an assessment of the database design and data distribution quality (Information Architecture Assessment) of key systems supporting the agency’s purpose or mission. These assessments support the TRM’s goal of optimizing service platform and infrastructure, component framework, and service interface and integration.

Data quality initiatives evaluate production databases based on certain quantitative and information-preserving transformation measures, such as data integrity, normalization, and performance. In large distributed systems, many decisions are made at design time based on the need for improved performance, but the tools for capturing a system’s performance measurement at run time (performance metadata) and for using this information to adaptively configure the system are often missing. Database performance assessments can provide those systems with the performance metadata they will need to become optimized after they have gone into production.

However, there are also many examples of database applications that are in most ways “well-formed” with high data quality and low data redundancy counts but lack semantic or cognitive fidelity (i.e., the right design). Whether the database meets the expectations of its end-users is only one aspect of overall database quality.

### 3.5.3 Align information architecture with data collection strategies

One of the major values of data quality initiatives is to minimize the data collection burden on customers and business partners. Data quality assessments generally weigh the efficiency and consistency of data collection/data storage in support of the TRM's goal of optimized component performance. The following performance areas can be measured and standardized across the enterprise:

- Consistency,
- Coverage/scope,
- Timeliness,
- Value in terms of cost,
- Accuracy/error rate,
- Accessibility,
- System performance/ease of use,
- Integration with other databases,
- Output,
- Documentation, and
- Customer support.

An information architecture assessment can reveal inefficiencies of data collection and processing, then documented in a manner that produces information consistent across agency lines of business. At the other end of the information product chain, developing consistent reporting protocols maximizes the utility of the information for identifying the successes of each individual line of business.

## 3.6 Conclusion

While this section of the DAS DQ Framework illustrates how federal agencies can leverage data quality principles and initiatives to aid in their FEA model development – particularly the development of the DRM as it is focused on data description and data sharing – Section 4 provides a structured framework for activating these industry-proven data quality initiatives into an agency-wide business program supporting business processes. The thirteen procedures explained in Section 4 provide federal agencies and COI's with repeatable processes for:

- Selecting and diagnosing data quality problems,
- Identifying important business processes necessitating DQI assessment,
- Applying standard data quality characteristics to detect faulty data,
- Establishing benchmarks or thresholds for quality,
- Certifying (statistically measuring) their quality, and
- Continuously monitoring their quality compliance.

## SECTION 4. IMPLEMENTING THE DATA QUALITY IMPROVEMENT (DQI) INITIATIVE

Reaping the substantial benefits associated with a DQI initiative requires a serious commitment on behalf of an agency. Its importance must therefore be effectively communicated to the business managers who will sponsor both the technology and the organizational infrastructure in order to ensure a successful program that will:

- Identify data quality problems,
- Correlate impacts to root causes,
- Calculate costs to remediate, and
- Project return on investment.

A thorough cost/benefit analysis is critical to any data quality program. It is the economic rationale for the added value brought by improved data quality that supports the multiple processes and steps required to implement the program. Executives in the agency will inevitably be reluctant to commit scarce resources to the effort unless there is a defensible justification to show the costs of neglecting the data quality problem and the economic benefits of improving the quality environment.

### 4.1 *Developing the DQI Business Plan*

Obtaining senior management support by means of a detailed Business Plan is the first step in selling the data quality value proposition to the agency. From executive-level appointments to the lower managerial strata, an understandable rationale must be given in order to obtain buy-in and motivate active participation in the data quality effort. Federal data quality projects will gain traction if executives institute incentive programs to encourage employees to follow the new data quality policies, and if the agencies publicly recognize employees who make major contributions toward the DQI process.

Following are the essential components of a detailed Business Plan.<sup>4</sup>

#### 4.1.1 **Strategic alignment perspective**

The Business Plan describes how the DQI initiative aligns with the organization's overall strategy or a component of the strategy. This should include a discussion of how the initiative also aligns with other existing improvement or quality programs. If the strategy does not specifically call out the value of the data or information as a strategic resource, then this strategic case should demonstrate how the improved data quality links into other outputs or outcomes described in the strategy as well as the vision and mission of the organization.

#### 4.1.2 **Alternative approaches**

The Business Plan provides a discussion of alternative approaches that may be applied as a DQI initiative. These can be choices between some of the data tools or combinations of those tools discussed in Section 5 of the DAS DQ Framework. A meaningful discussion of alternative approaches recognizes that options are available and provides a better basis for a decision to support any recommended approach. Some of the areas that can be addressed in reviewing alternative approaches include:

---

<sup>4</sup>Paul Harmon, Business Process Trends (resources), [www.bptrends.com](http://www.bptrends.com) and National Defense University, Information Resources Management College *Strategies for Process Improvement* Course materials, available at [http://www.ndu.edu/IRMC/interactive\\_schedule/course\\_descriptions/pri-details.html](http://www.ndu.edu/IRMC/interactive_schedule/course_descriptions/pri-details.html)

- Alignment with current quality management systems and other improvement processes in the organization,
- Ease of implementation,
- Sustainability potential (i.e., survivability potential when there is a change in leadership),
- Alignment with established performance measurement systems , and
- Resource needs.

### **4.1.3 Performance improvement perspective**

Discussing how the organization's performance improves directly with improved data quality demonstrates to managers and others interested in the organization's performance how the DQI initiative may have measurable value for the enterprise. The various data improvement tools described in Section 5 suggest potential performance measures commensurate with the tool selected. It is acceptable to select one or more performance measures and one or more DQI activities that are directly applicable to each performance measurement.

### **4.1.4 Project management perspective**

The Business Plan should include a description of the methodology that will be used to manage the DQI process as an ongoing project. This can include how the initiative will co-exist with other program management methodologies including how progress will be tracked and reported. One useful exercise is to map any elements of the chosen DQI initiative to project management elements.

### **4.1.5 Financial perspective**

A discussion of the anticipated costs and savings expected is appropriate for a DQI initiative. This can be a particularly powerful section of the Business Plan if the financial perspective addresses the potential for resources savings accomplished through new anticipated efficiencies or the elimination of waste, such as scrap and re-work in making corrections to faulty data.

### **4.1.6 Information perspective**

The Business Plan cannot simply address how data quality will be improved; it must also include how and if the information management systems of the organization may be impacted and what new detailed data and functional requirements may need to be addressed.

### **4.1.7 Change management approach**

In addition to measures that show how the DQI initiative will improve the organization's performance, the Business Plan should address how information will be collected and utilized to track and report progress on the success of the effort. Managers need information to manage the change process and staff need feedback in the form of specific measures. Individual DQI processes should be measured and tracked.

The change management processes explicitly track some or all of the performance measures identified previously (see Section 4.1.3 above). Tracking these performance measures ensures that objectives are met and that appropriate course corrections are implemented if expected performance improvements do not occur and/or new issues emerge.

### 4.1.8 Next steps

The final step in the Business Plan is to clearly lay out the DQI activities comprising a written Data Quality Plan for the project, which might be enterprise in scope or more local to a specific business area. The Data Quality Plan should include the following:

- Senior managers' recognition and endorsement of DQI.
- Designation of a management champion to own the process and ensure alignment with quality programs and other improvement activities.
- Establishment of a group to work interactively with representatives throughout the federal program offices.
- Commitment of resources.
- Selection of DQI activities in a phased, carefully scheduled approach.

Sections 4.2-4.14 discuss the major activities that may comprise a successful Data Quality Plan:

- Identify Data Quality Scope
- Conduct Root Cause Analysis
- Perform Information Value Cost Chain (VCC) Analysis
- Set Data Quality Metrics and Standards
- Assess Data Against Data Quality Metrics
- Assess Information Architecture and Data Definition Quality
- Evaluate Costs of Non-Quality Information
- Develop DQ Governance, Data Stewardship Roles
- Assess Presence of Statistical Process Control (SPC)
- Implement Improvements and Data Corrections
- Develop Plan for Continued Data Quality Assurance
- Educate the Government Culture
- Save Data Quality Products to Enterprise Metadata Repository

These activities or process areas are arranged in Figure 4-1 below into an interactive template constituting three major activity levels, against which federal agencies may benchmark the extent of their own data quality practice. In the graphic, blue-shaded processes are chiefly at the enterprise level, yielding maximum return on investment (ROI) if the majority of DQI is centered among these activities in a state of continuous quality maintenance and improvement. Among the players involved in data quality, these activities are generally the responsibility of Executive Management working in concert with the Data Quality Group. Gray-shaded activities should chiefly be the responsibility of the Data Administrators (medium-to-high ROI), while red-shaded activities, the chief responsibility of Metadata Administrators, are information system-level processes yielding the least ROI if conducted solely by themselves.

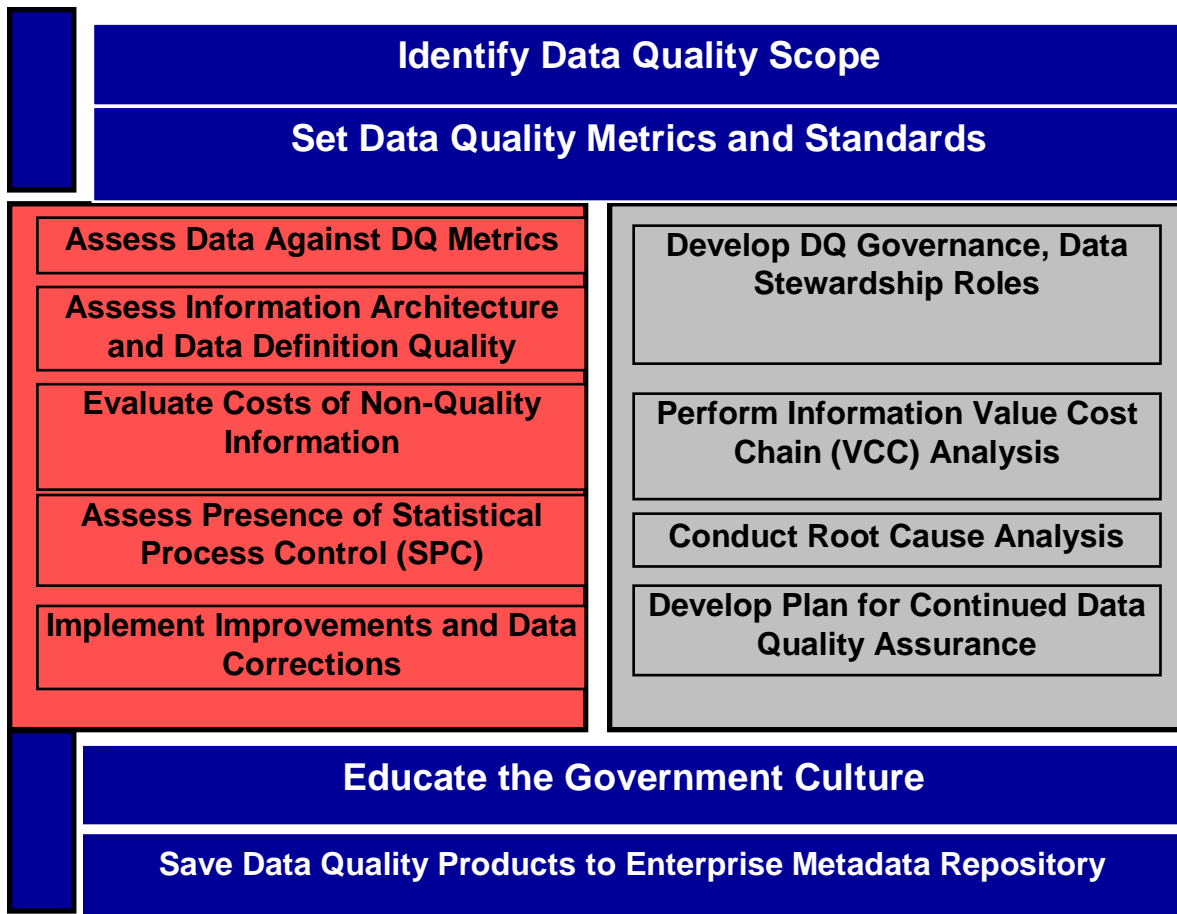


Figure 4-1: DQI Process Framework

## 4.2 Identify Data Quality Scope

Although all federal agencies desire cleaner data, the specific, strategic goal for obtaining higher quality will differ for each. Typically, data quality projects are selected by information system administrators and/or those who use the data. For example, users may report frustration with errors in the data recorded in system tables and/or records. Proper system controls may not have been implemented to minimize the possibility of quality problems, whether through data entry or external/internal system feeds. The list of possible problems is endless, but data quality issues tend to fall into the following Error Types:

1. Data-Centric Problems
2. Training Problems
3. Policy Problems
4. Procedure Problems
5. Internal System Problems
6. Interface Problems
7. Other (cultural or environmental)

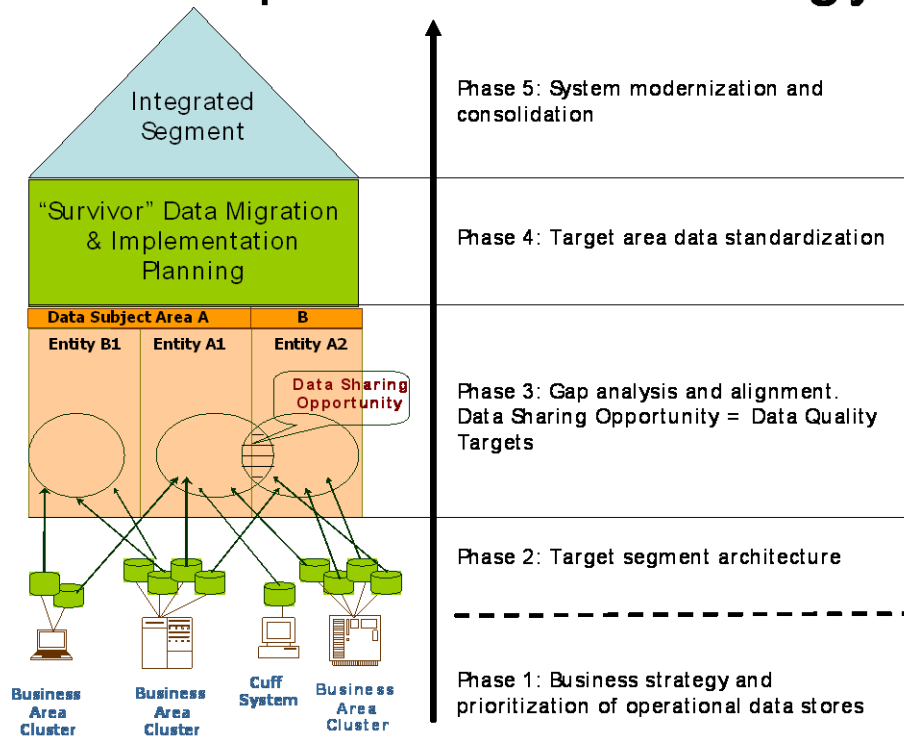
Data quality issues of Error Types 2-7 can usually be solved by traditional means (see Table 4-1 below for recommended actions concerning these issues). For example, while a Training problem may require developing new training programs and an Interface problem may involve improvement in

system-to-system design, they are nevertheless data quality problems: they all have to do with impediments to the integrity, accessibility, usability or interpretability of mission-critical information being processed by the organization.

Data-centric quality problems of Error Type 1, on the other hand, are issues surrounding persistent data and their behavior and can be much more difficult to solve. Data-centric problems have to do with the collection, storage and retrieval of data – the entire lifecycle of the data’s transaction management – and require the kind of structured DQI-solutions approach articulated in Sections 4.4-4.12 of this document.

An example of a data-centric quality driver may come as a result of an agency’s data architecture compliance. In this regard, an agency may already have mapped enterprise information flows and developed target, segment data architectures based upon the mappings. It may therefore be prudent and cost-effective to commence DQI to run in tandem with the segment data architecture development, segment by segment. The following graphic (Figure 4-2) is a generic example of the phases of segment data architecture development currently used at the federal level and the opportunities for DQI within the sequence.

# Segment Data Architecture Development Methodology



**Figure 4-2: Segment Architecture Development.** *Mapping systems data stores to Data Subject Areas and Entities can discover common needs.*

In the development sequence, the Venn Diagram produced during Phase 3 is marked as a “Data Sharing Opportunity.” This intersection of business entities yields a probable scope of data sets that can be inspected and monitored for quality.



The need to comply with internal or external regulations is another data-centric issue that may require attention. An example of an internal regulation might be an audit by the agency's Inspector General specifying a data quality problem (e.g., the data underlying Annual Performance Plan reporting has been found to be inaccurate). An example of an external regulation might be the OMB's Federal Funding Accountability and Transparency Act (FFATA), for which federal agency's must have on file and in a searchable format certain data elements that capture the key information required by the FFATA memorandum. These FFATA data elements would be immediate candidates for data quality inspection and also specify a requirement for monitoring any new information systems that fund grants, loans, awards, cooperative agreements, or other forms of financial assistance to American citizens.

Lacking these drivers, the following sequence may be used to determine mission-critical data assets that can be targeted for a formal DQI process where a number of diverse data-centric quality problems are known to exist:

- Conduct focus groups and/or distribute data quality surveys across every program area to elicit data quality problems and determine the agency's strategic goals,
- Connect data quality problems to strategic goals in order to prioritize key data for improvement,
- Develop a "Data Quality Plan" or roadmap for improvement,
- Select data (databases, files, other storage mechanisms) for assessment based upon the Data Quality Plan,
- Determine appropriate timeframes to conduct the assessment,
- Perform a preliminary feasibility study of data repositories in scope to determine if complete baseline data (preferably three years in age or older) are available for inspection, and
- Make a "Go/No Go" decision on the assessment of data and the systems that support the data.

The steps above will be performed initially to assess the data quality of mission critical data within the immediate scope and, subsequently, to ensure the target compliance levels are maintained. The long-term goal is to achieve a state of continuous improvement, yielding agency-wide data of the highest quality.

### 4.3 Conduct Root Cause Analysis

Once the data quality scope has been determined, it is necessary to conduct a formal Root Cause Analysis as the immediate next step. This analysis will seek to identify one or more DQ Error Types and Focus Areas for each DQ Issue uncovered (see Table 4-1 below). Most error types fall into one of the first six categories, while the last category is utilized for unusual errors.

<i><b>DQ ERROR TYPE</b></i>	<i><b>FOCUS AREA</b></i>	<i><b>RECOMMENDED ACTION</b></i>
1. <b>Data-centric Problem</b> – The problem is with persistent data itself: how it is collected, distributed, stored, and managed. The data do not conform to their intended business rules and business purpose.	A. Data Managers & Designers – relates to some misalignment between data collection, storage, and appropriate use.	Commence a formal DQI data assessment/ improvement process articulated in Section 4.4 – 4.12 of this document.
2. <b>Training Problem</b> - Human impact problems regarding knowledge of established and or adequate policy/procedures. Requires training improvement.	B. Collectors - relates to the input of data	Improve/develop/administer training courses related to the data issue.

3. <b>Policy Problem</b> - Policy not yet established, policy that needs revising, or a failure on the part of knowledge workers or managers to comply with one or more policies. Requires policy and procedural improvements.	C. Custodians - relates to the maintenance of the data	Resolve conflicts in existing policies and procedures and develop appropriate guidance that will institutionalize the behaviors that promote good quality.
4. <b>Procedure Problem</b> - Procedures not yet established, procedures that need revising, or a failure by knowledge workers or management to comply with written or implied procedures. Requires process improvement.	D. Customers - relates to the users of the data	Improve the functional processes that are used to create, maintain and disseminate the data.
5. <b>Internal System Error</b> - Errors that are resident in the data system automated programming code. This includes problems caused by insufficient edits, software and hardware. Requires system improvement.	E. Components – relates to applications and technology.	Software, hardware and telecommunication changes can aid in improving data quality.
6. <b>Interface System Error</b> - Data errors occurring when two or more data systems share data values. Requires interface improvement.	F. Architecture – relates to system-to-system design and data exchange factors.	Improve system to system interfaces as well as overall data design within the systems.
7. <b>Other Errors</b> - All errors that do not fit into above categories, including an unwillingness to accept change and promote necessary data quality improvements.	G. Other. May be environmental or cultural factors.	Requires extraordinary measures other than those previously mentioned.

Table 4-1: Root Cause Analysis Worksheet

#### 4.4 Perform Information Value Cost Chain (VCC) Analysis

Data that have been identified as comprising the scope of a data-centric quality problem are then organized into Information Groups (also referred to as Information Products or “IP”), each having a common business purpose managed by distinct sets of information stakeholders. The VCC facilitates the identification and full understanding of this data across the enterprise.

There are several ways to diagram an information value chain, and one of the most common ways is to create an Information Product Map (“IP Map”). A sample IP Map is presented in Figure 4-3, showing the life cycle movement of a single data element (‘inspection\_code’) integral to a physical inspection business process. The IP Map:

- Identifies the files/databases which include the IP’s attributes (i.e., data elements),
- Identifies the database of origin and database of record (see Glossary for definitions of these terms),
- Identifies external sources of the data,
- Illustrates the movement of IP attributes between files/databases,
- Identifies interface points between systems where data are either duplicated or transformed,
- Facilitates the identification of stakeholders, and
- Can be leveraged for analysis of other IPs within a file/database.

Inspection\_code is initially created by an Inspector on his palm device, shown in the upper left hand corner of Figure 4-3. The sub-routine depicted in the diagram is a sequence in which “no inspection violation” has occurred during the physical inspection (shown as a diamond labeled ‘D56’ in the middle of the diagram, resulting in a component data transfer ‘CD64’). Quality Block ‘QB65’ checks the true value of inspection\_code against a look-up table and tests its form and content against applicable data quality measurements (see Section 4.5.1). Assuming that inspection\_code passes the consistency test, it is processed by a Warehouse Report Generator and included in the finished Inspection Summary information product. The database of origin for this data element is represented by a shaded cylinder, in this case STO54 SQL, while the database of record is a cylinder outlined with a heavy, bold line (STO89 SAS Data Warehouse). Transformation and aggregation rules are communicated by the structure of the lines themselves. Additional information can be collected for each data flow to make it more granular, such as data domain descriptions (through captioning of the data flow arrows) and cost estimates for storage and/or transmission at each stage (also through captioning).

### Physical Inspection Information Product (IP) Map

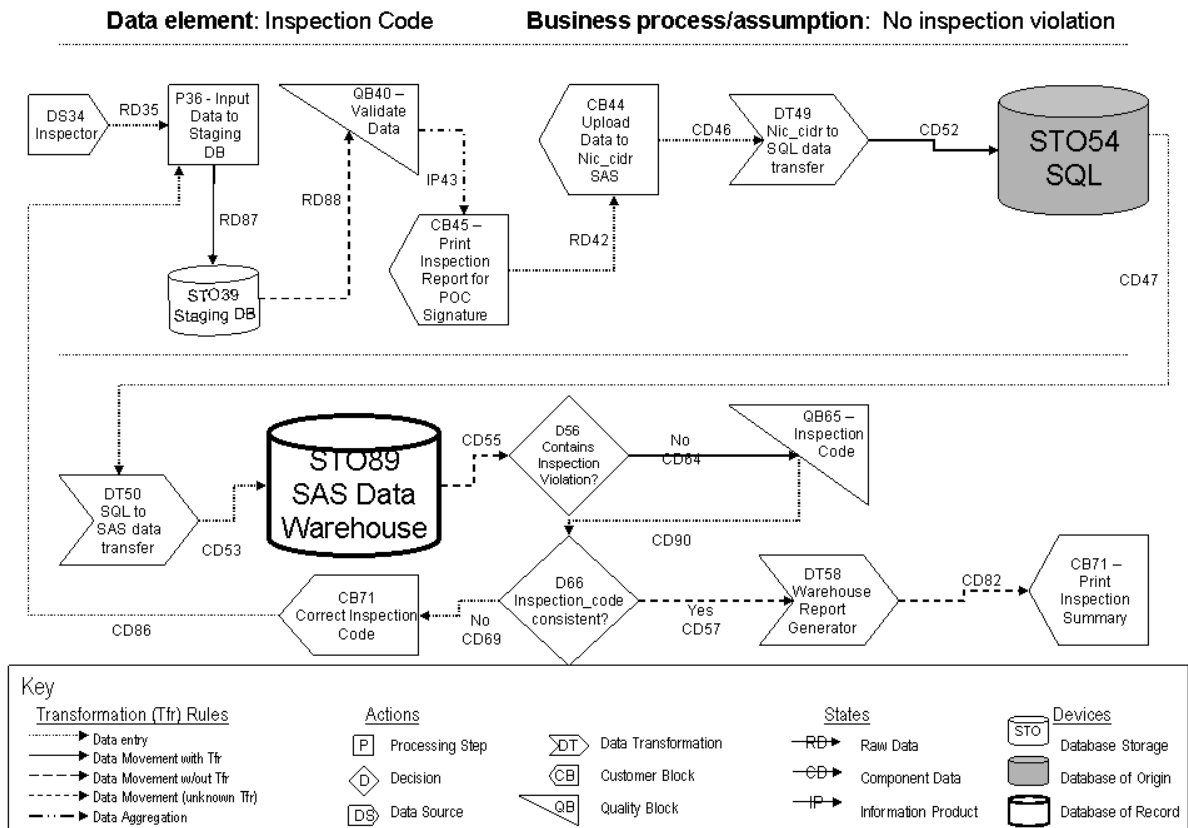


Figure 4-3: Sample Information Product (IP) Map

## 4.5 Set Data Quality Metrics and Standards

Data quality metrics ordinarily reflect the explicit as well as the implicit business principles of an agency. Business principles are explicit if stated in formal documents such as mission or vision statements, implicit if they are not. For example, if an agency rewards project managers for meeting deadlines even though their applications are full of errors, while it punishes project managers for missing deadlines even though their applications are flawless, then the implicit principle is “speed before quality.” Therefore, when creating data quality metrics the explicit as well as implicit business

principles must be reviewed and changed, if necessary, to support the metrics. To better understand the agency's business principles, it may be necessary to conduct interviews with key business personnel to find out how they and their information stakeholders outside of the organization are using the data, to learn which data quality dimensions are most important, and to determine these dimensions' quality expectations.

Another important aspect to measuring data quality is setting goals. Agencies need to be clear on where they are today and what they are trying to achieve in the short term, medium term, and long term. What are the agency's priorities? Should operational data be addressed or only analytical data? Should financial data be cleansed first or a specific subject area for an application, such as customer relationship management (CRM)? What is the plan for incrementally managing data quality improvements? What are the staffing requirements and what are the roles and responsibilities for a data quality improvement initiative? These questions must be answered to develop meaningful and actionable data quality metrics.

#### 4.5.1 Key data quality measurements

Assessment of data quality requires assessments along a number of dimensions. Each agency must determine which of the following dimensions in common use are most important to its operations and strategic goals. For example, in many cases it may be unfeasible to conduct Accuracy assessments due to the sheer expense and effort of comparing actual values (from actual forms, instruments, or other collection media) with the data representations collected. In addition many of these variables are context-dependent: although falling within a specific dimensional category, the specific measure to assess a specific dimension will vary from agency to agency.

There are two sets of dimensions: data content and presentation. The data content quality dimensions are described in Table 4-2.

<b>Dimension</b>	<b>Quality Dimension Description</b>	<b>Example of Non-Quality Data</b>
Validity	The degree to which the data conforms to its definition, domain values and business rules.	A U.S. address has a state abbreviation that is not a valid abbreviation (not in the valid state abbreviation list).
Non-Duplication	The degree to which there are no redundant occurrences or records of the same real world object or event.	One applicant has multiple applicant records (evident when an applicant gets duplicate, even conflicting, notices).
Completeness	The degree to which all required data is known. This includes having all required data elements (all facts about the object or event), having all required records, and having all required values.	An indicator for spouse is set to "yes", but spousal data is not present.
Relationship Validity	The degree to which related data conforms to the associative business rules.	A property address shows a Michigan zip code, but a Florida city and state.
Consistency	The degree to which redundant facts are equivalent across two or more databases in which the facts are maintained.	The same applicant is present in two databases or systems and has different name, address, or dependents.
Concurrency	The timing of updates to ensure that duplicate data stored in redundant files is equivalent. This is a measure of the information float (the time elapsed from the initial acquisition of the information in one file or table to the time it is propagated to another file or table).	On Monday, an applicant's change of address is updated in the Applicant record of origin file, but the record is propagated to the main Program database after the weekend cycle (Friday night). That record has a concurrency float of 5 days between the record-of-origin file and the record-of-reference database.

<b>Dimension</b>	<b>Quality Dimension Description</b>	<b>Example of Non-Quality Data</b>
Timeliness	The degree to which data is available to support a given information consumer or process when required.	A change of address is needed to schedule an inspection but is not available to the field office, and the inspector leaves without the proper information.
Accurate (to reality)	The degree to which data accurately reflects the real-world object or event being described.	The home telephone number for a customer record does not match the actual telephone number.
Accurate (to surrogate source)	The degree to which the data matches the original source of data, such as a form, application, or other document	An applicant's reported income on the application form does not match what is in the database.
Precision	The degree to which data is known to the right level of detail (e.g., the right number of decimal digits to the right of the decimal point).	The summary amounts in congressional reports are rounded to the nearest \$1,000.00 and do not include amounts in the hundreds, tens, dollars or pennies. However, the amounts will be aggregated in dollars and cents and then rounded to the nearest \$1,000 to avoid rounding errors.
Derivation Integrity	The correctness with which derived data is calculated from its base data.	The summary of accounts for a given district does not contain all valid entries for the district.

**Table 4-2: Dimensions of Data Content Quality**

The presentation quality dimensions are listed in Table 4-3.

<b>Dimension</b>	<b>Quality Dimension Description</b>	<b>Example of Non-Quality Data</b>
Accessibility	A measurement of the degree of ease-of-access interested information consumers have to the data they require.	The planning analyst needs the current account of insurance per jurisdiction, but the information is not available unless a programmer extracts it.
Contextual Clarity	The degree to which presentation of the data enables the information consumer to understand the meaning of the data and avoid misinterpretation (intuitiveness).	Applicants report incorrect, or have missing, annual income on the form due to an improper label.
Usability	The degree to which the information presentation is directly and efficiently usable for its purpose.	Statistical information that would be easily understood if presented in a table format is provided in several paragraphs of text.
Rightness	The characteristic of having the right kind of data with the right quality to support a given process.	All the application information is present, but the credit report is missing, so the underwriting process cannot be executed.

**Table 4-3: Dimensions of Data Presentation Quality**

Setting the proper standards for each dimension is based on the agency's desired quality class of the data. The quality classes defined below indicate the degree of quality required for the particular data under consideration, based on business need. Following each quality class definition, a recommended "sigma level" is provided. The Six Sigma methodology, a popular variant of Total Quality Management, has defined widely-accepted, standard, quality-level benchmarks. The three data quality classes are:

- **Absolute** (zero-defect or close to zero-defect) indicates this data can cause significant process failure when containing defects. The recommended quality standard is 6 sigma, or no more than 3.4 errors per million opportunities (i.e., total database records sampled).
- **Second Tier** (high cost of non-quality) indicates there are high costs associated with defects in these data, and, therefore, it is critical to keep defects to a minimum. Achieving this standard for data in this class would be possible through ongoing monitoring and correcting of data (statistical process control), which is a more cost-effective approach than engineering near zero-defect application edits for all non-compliant system data. The standard of 4 sigma (6,210 errors per million, or a 0.62% error rate) is appropriate for this class, since it represents industry-standard, real-world production requirements.
- **Third Tier** (moderate cost of non-quality) indicates the costs associated with defects in this data are moderate and must be avoided whenever possible. Quality tolerance for this level should be no worse than 3 sigma, or 66,807 errors per million.

#### 4.6 Assess Data Against Data Quality Metrics

Whereas the DQI process step “Set Data Quality Metrics and Standards” answers the “what?” this step answers the “how?” A data-centric improvement cycle includes an assessment of the data in scope against the data quality standards for each dimension defined in Section 4.5 above. This can either be an initial enterprise-wide data quality assessment, a system-by-system data quality assessment, or a department-by-department data quality assessment. Another type of assessment is a periodic data audit. This type of assessment is usually limited to one file or one database at a time. It involves data profiling as well as manual validation of data values against the documented data domains (valid data values). These domains should have already been documented in the metadata, but if not, they can be found in programs, code translation books, online help screens, spreadsheets, and other documents. In the worst case, they can be discovered by asking subject matter experts.

When performing the assessment, the Data Quality Group should not limit its efforts to merely profiling the data, performing Data Content Quality testing against the identified business rules, and collecting statistics on data defects. The entire data entry or data manipulation process must be analyzed to find the root causes of errors and to find process improvement opportunities.

Depending upon the assessment objectives, data may need to be measured at different points in the VCC diagram. If budget concerns and time constraints permit, it is preferable to pinpoint the system that initially captures the data in scope (database of origin) and assess the same data in all applications and files and/or other databases along the information supply chain. Within government agencies, there is a tendency to assess the quality only in the circle of influence (the owned application or database); however, the critical impact to an agency occurs when the data are not of the expected quality but is shared across other applications and business areas. If there are resource or time constraints, it is better to reduce the number of data elements in the assessment but include the entire value chain for the data being assessed. This means that the assessment must include all relevant databases, applications, files, and interfaces. In all cases, the approach to be taken must be defined and documented.

Due to time and resource constraints, it may be possible to measure data in only one location, when there are many other systems handling the data during the data’s life cycle. In this case, it is important to verify – through careful inspection of ALL data upload/transfer programs along the entire VCC – that the data has integrity and has not been filtered or corrupted in any way.

Assessment Objective	Assessment Point
1. Understand state of quality in the database.	The entire database or file. This should be a data source that supports major business processes.

<b>Assessment Objective</b>	<b>Assessment Point</b>
2. Ensure effectiveness of a specific process.	The records output from the processes within a time period being assessed but prior to any corrective actions.
3. Identify data requiring correction.	The entire database or file. This should be a data source that supports major business processes.
4. Identify processes requiring improvement.	The records output from the processes within a time period being assessed, but prior to any corrective actions.
5. Ensure concurrency of data in multiple locations.	A sample of records from the record of origin that must be compared against equivalent records in the downstream database. If data may be created in the downstream database, records are extracted from one to find the equivalent records in the other.
6. Ensure timeliness of data.	A sample of data at the point of origin. These must be compared against equivalent data from the database from which timely access is required.
7. Ensure effectiveness of data warehouse conditioning process.	A sample of data from the record-of-reference. These must be compared against equivalent record(s) in the data warehouse.

**Table 4-4: Data Quality Assessment Point by Assessment Objective**

## **4.7 Assess Information Architecture and Data Definition Quality**

An information system should be engineered in such a way that the data collected and managed by the system aligns with its business strategy as well as the information architecture. This facet of data quality – information architecture and data definition best practices – starts with identifying the data sets used by the application and enumerating the data attributes within each data set. Each data element must have a name, a structural format, and a definition, which must be documented within a core metadata repository. Each data set models a relevant business concept, and each data element provides insight into that business concept – all within the context of the business area “owning” the application. In turn, each definition must be subjected to review to ensure that it is correct, defensible, and is grounded by an authoritative source. This implies that the agency have a data governance structure in place to perform these reviews, as well as metadata management policies and procedures. Every activity that creates, modifies, or retires a data element must somehow support a business activity contributing to the business area’s overall business objectives.

When assessing the information architecture, one must document each information function and how it maps to achieving business objectives. A standardized approach for functional description will help in assessing functional overlap, which may be subject for review as the core master data objects are identified and consolidated. However, in all situations the application functionality essentially represents the ways that information policies are implemented across the enterprise.

Poor data definition and information architecture quality undermines an agency’s ability to create, maintain, and exploit quality data. Therefore, to support the objectives of the FEA Technical Reference Model, an agency must provide policy to support all business areas with a solid foundation for the development and maintenance of databases and files and the consistent management of the data assets across the enterprise.

### **4.7.1 Information architecture assessment**

Assessing the information architecture leads to important findings regarding the quality of the mechanism an information system employs for ensuring that data are well managed within the environment and distributed in an accurate, reliable format within the system’s repository and to other

units within the organization based on business need. A well designed information architecture allows disparate data to be captured and funneled into information that the business can interpret consistently for reporting past results and planning appropriately for the future. Inadequately designed information architecture is out of sync with the functional requirements of the business area – or is not scaleable enough to adapt to changing requirements – leading to a misalignment between the technical implementation of the information flow and the demands made to use the information for statistical reporting. This misalignment can impact the system’s data quality findings.

An information architecture assessment includes an inspection of the database’s logical data model, the database design (implementation model), and the physical implementation of the data structures against modeling, design, and implementation best practices. The assessment will determine: Whether the data model has all required entity types and attributes to support the business processes.

- Whether the data model truly reflects the real world entity types, attributes, and relationships.
- Which instances of data redundancy in proprietary files are controlled and which are not controlled (i.e., has denormalization of the original hierarchy been done for sound performance reasons?).
- Whether data are being captured as close to the original sources as possible.
- Whether new data products are being created “just in time” (minimizing the need for changes due to normal churn).
- Whether there is adequate exception handling (error catching).

#### 4.7.2 Data definition quality assessment

Data definitions must be consistent across the enterprise. The VCC diagrams developed by tracing in-scope data’s life cycle identify the files and databases where each data element is stored. A comparison of the data definitions and storage format definitions across these files and databases should be made to determine the level of consistency across the enterprise.

In cases where no formal data definitions have been compiled or maintained, it is still possible to derive definitions through “data profiling”. Data profiling is the measurement and analysis of the attributes of a data set using direct observation. Data profiling may include:

- Domain and validity analysis (forensic analysis).
- Identification of possible primary and foreign keys.
- Analysis of the database loading program (software code) for rules by which data columns are generated.
- Observation of the number and types of defects in the data such as blank fields, blank records, nulls, or domain outliers, tested against the preliminary rules developed during the forensic analysis.

If no data definitions are available and if data profiling does not yield conclusive, harmonized domain or validity rules, a focus group or other general discussion forum with all interested stakeholders must develop a general enterprise definition to be used for data quality assessment and other business purposes, and eventually for standard use across the enterprise.

### 4.8 Evaluate Costs of Non-Quality Information

After opportunities for improvement have been defined, the improvements should be analyzed, prioritized, approved and funded. Not all improvements have the same payback and not all improvements are practical or even feasible. An impact analysis should determine which improvements have the most far-reaching benefits. After improvement projects have been prioritized,



approved and funded, they should be staffed and scheduled. A template for determining non-quality information costs is included in Table 4-5 below. These benefits fall under the category of “tangible” benefits.

Non-Quality Information Costs				
Direct Costs Worksheet				
Information: _____	Cost per Instance	Number of Instances	Total Number per Year	Total Cost per Year
Process: _____				
Time: (loaded rate / hour = _____ / Hour)				
-				
Money				
-				
Materials				
-				
Facilities and Equipment				
-				
Computing Resource				
-				
Total Annual Costs				

**Table 4-5: Non-Quality Information Costs Worksheet**

Improving the quality of data within an agency can also result in “intangible” benefits. Intangible benefits are difficult to measure due to their qualitative and subjective nature. These include improved speed to solutions, improved customer satisfaction, improved morale, and consistency between systems. Due to the fact that most data quality benefits are considered intangible, the Direct Costs Worksheet process may not be applicable for all systems and products.

#### **4.9 Develop Data Quality Governance, Data Stewardship Roles**

Please refer to Section 3.2.1 of this document for a discussion of best practices in organizing the appropriate governance structure for data quality improvement, as well as the supporting groups who will sustain its practice.

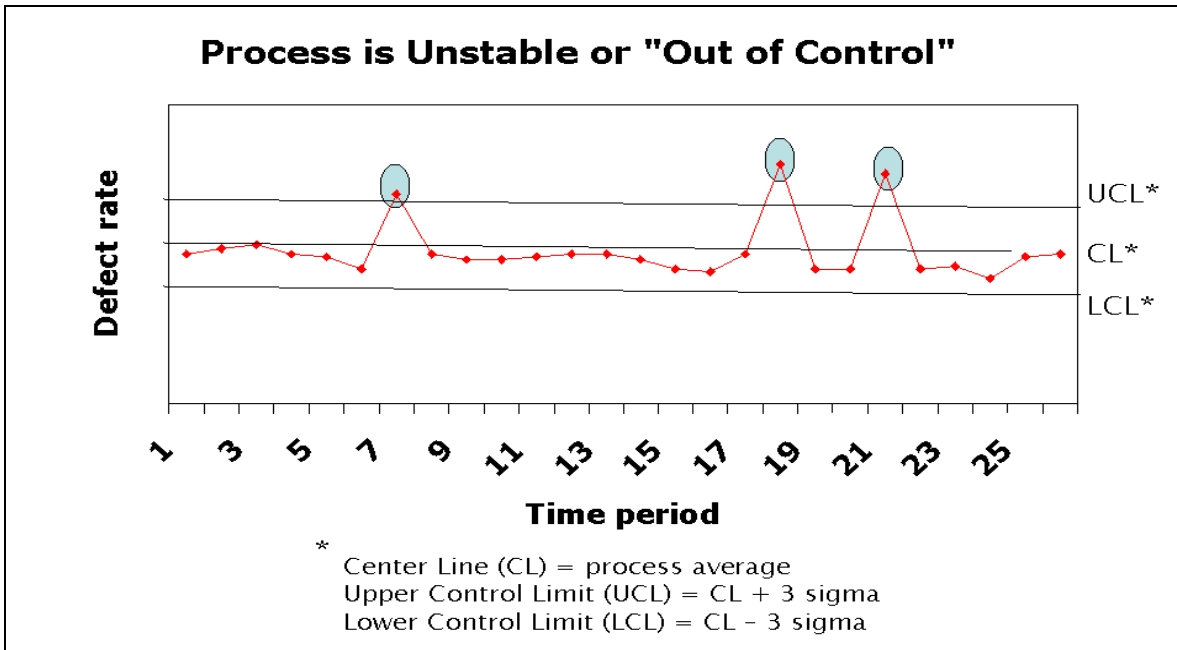
#### **4.10 Assess Presence of Statistical Process Control (SPC)**

For processes that acquire, produce or maintain federal data, best practices recommend that the data be in a state of statistical control. A comprehensive DQI program at the federal level assesses the presence or absence of Statistical Process Control (SPC) used among transactional systems throughout the enterprise. SPC is something that federal program areas can conduct themselves if there is a limited budget for DQI, without the direct involvement of the Data Quality Group.

To enable the objectives of the FEA Service Component Reference Model, federal business areas should continually monitor mission-critical data quality using appropriate sampling, measuring and tracking to measure progress toward the quality standard established for the data. The processes embedded in SPC measure the accuracy of critical data, establish performance benchmarks, and quantifiably evaluate data as they are being collected. The benchmarks evaluate variances inherent in collection protocols that may impact the repeatability or consistency of future data collection. SPC requires that data collected through business objective should periodically be aggregated and analyzed to establish a standard level of concurrence (process mean). Analyses are then conducted to

identify variances within various inspectable levels (common causes of variation). Standards for data accuracy are determined by the level of concurrence attained by the N<sup>th</sup> percentile of the data collection activities satisfying the business objective and as measured during the agreed-upon period of inspection (desired level of confidence). This standard is used to determine the acceptability of the activities as measured over the course of the next months (variability over time). Data collection activities that do not meet or surpass the standard (special causes) are to be rejected and the activities commenced again and re-processed. The information gathered during this evaluation is used to make improvements to the protocol surrounding the business objective.

The following graphic (Figure 4-4) shows a typical “run chart” that might represent a time series plot of data variance.



**Figure 4-4: SPC Run Chart**

In the run chart above, data collected during the 7<sup>th</sup>, 18<sup>th</sup> and 21<sup>st</sup> time periods show abnormal performance exceeding the Upper Control Limits previously established. This unstable or “out of control” process demands that some sort of adjustments to the process – which may include training initiatives, modifications to the data collection activities, or modifications to other protocols involved in the business objective – need to be made (action to bring process back within statistical control).

#### **4.11 Implement Improvements and Data Corrections**

The principles of continuous data quality improvement analyze the root causes of defective data and implement improvements that manage data for quality throughout their life cycle. Eliminating the causes of data defects and the production of defective data builds quality in and reduces the need to conduct data correction activities.

In most cases, the Data Quality Group can implement the approved improvements, but in many cases, other staff members from both the business side and IT will be required. For example, a decision might have been made that an overloaded column (a column containing data values describing multiple attributes) should be separated in a database. That would involve the business people who are currently accessing the database, the database administrators who are maintaining it, and the developers whose programs are accessing it.

Improvement consists of selecting the process for data quality, developing a plan for improvement, implementing the improvement in a controlled environment, checking the impact of the improvement to make sure that results are as expected, and standardizing the improvement across the enterprise. Unlike data quality improvement, which is a continuing effort, data correction should be considered a one time only activity. Because data can be corrupted with new defects by a faulty process, it is necessary to implement improvements to the data quality process simultaneously with the data correction.

Improvements can be a mixture of automated and manual techniques, of short, simple implementations and lengthy, complex implementations that are applied at different times. Because of the possible diversity of improvements, agencies must track progress closely. Documenting the successes and challenges of implementation allows sharing and re-use of the more effective DQI techniques.

The implementation of DQI will include one or more of the following actions:

- The implementation of awareness (education) activities (see Section 4.13).
- The implementation of statistical procedures to bring processes into control.
- Improvements to training, skills development (including mentoring) and staffing levels.
- Improvements to procedures and work standards.
- Changes to automated systems and databases.

Although some impact analysis will have been performed during planning, occasionally an adverse impact will be overlooked. Or worse, the implemented improvement might have inadvertently created a new problem. It is therefore advisable to monitor the implemented improvements and evaluate their effectiveness. If deemed necessary, an improvement can be reversed.

## Repeated Application of the Data Quality Improvement (DQI) Process

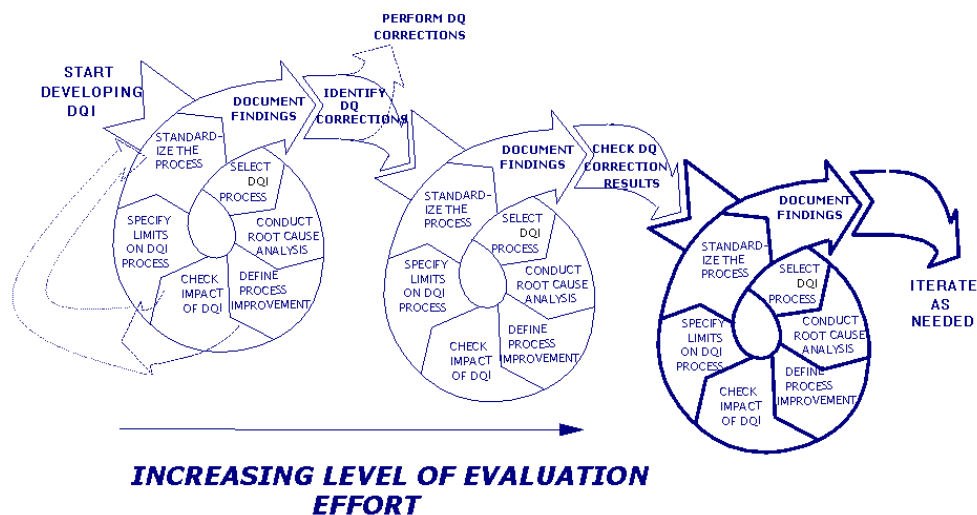


Figure 4-5: Iterative Data Quality Improvement (DQI) Process

## **4.12 Develop Plan for Continued Data Quality Assurance**

A Data Quality Assurance Plan documents the planning, implementation, and assessment procedures for maintaining continuous DQI, as well as any specific quality assurance and quality control activities. It integrates all the technical and non-technical aspects of the DQI in order to provide a "blueprint" for obtaining the type and quality of data needed for specific business decisions or uses.

A quality assurance strategy, whether for data quality or other products or services, must contain certain key characteristics that tie it to the quality of a business area's program outputs and outcomes.

- A written definition or specification that defines what is to be measured, including a clear description of the standard for acceptance (for example, data will be accurate, consistent, and timely, without errors or inconsistencies and no older than two years).
- A quality control specification which defines the procedures for measuring, evaluating, and controlling for various characteristics, including:
  - identifying the key activities in each process which have a significant influence on characteristics (for example, data collected from outside sources),
  - analyzing the key activities to select those characteristics whose measurement and control will ensure quality (for example, outside source reporting instructions),
  - defining the methods for measuring and evaluating the selected characteristics (for example, conducting oversight and verification visits to outside data providers), and
  - establishing the means of controlling the characteristics within specified limits (for example, written instructions to outside data providers coupled with a sampling strategy of cases or files).
- Clear accountability and authority for the quality assurance role within the business area, and the role of performance measurement within the accountability process.
- A review strategy that periodically selects and measures statistically valid samples for a quality review (for example, sampling would be done on a monthly basis).
- A feedback mechanism within the system that measures, communicates, and corrects instances of non-compliance with quality standards (for example, a written summary of the results of sampling and oversight visits would be furnished to all parties involved showing results and needed areas of improvement).
- Third-party calibration/verification of the measurements and techniques being used and of the inspection results (for example, the Inspector General or Chief Financial Officer would include in its annual audit plan data quality reviews).

## **4.13 Educate the Government Culture**

One of the most important steps in the DQI process is to disseminate information about the new improvement process just implemented. Depending on the scope of the change, education can be accomplished through classroom training, computer-based training, an announcement on the agency's Intranet, an internal newsletter, or simple e-mail notification.

Data quality training should be instituted at the business office (or system) level to address poor data entry habits. Not all data rules can be enforced through edit checks, data defect prevention tools, or by the features of relational databases, such as strong data typing, referential integrity, use of look-up tables, and the use of stored edit procedures. Many data violations can still occur because of human error, negligence, or intentionally introduced errors. For example, if an end user needs a new data element but must wait six months for IT to change the database, then the end user might simply decide to overload an existing column and use it for dual (or triple) purposes, such as putting the date of the last promotion into the Account Closed Date column.

A clear statement of policy must be in place for government entities to remain engaged and to succeed in maintaining a viable, continuous data quality effort, which in turn proactively supports business activities. The data quality policy can be articulated in the form of a Data Quality Handbook, which should be written in such a way that not only technical but also business personnel can understand it.

A complete data quality handbook has different levels of detail ranging from broad guidelines, roles and responsibilities, to more detailed elaboration and specification of these guidelines. It should address both data quality practice – such as management issues, implementation, operational issues and standards – as well as the data product itself. Federal agencies that take their Data Quality Handbook through clearance (i.e., the Handbook becomes official policy) are stating their intention to maintain data quality as part of the agencies' business agenda. In this case, the Handbook will include sections documenting the role of data in its business strategy and operations, and continually making the case why an enforceable data quality program is necessary to the agencies' ongoing business mission.

#### **4.14 Save Data Quality Products to Enterprise Metadata Repository**

The best data quality program in the world ultimately will fail if the results of the DQI are not saved to an enterprise-level repository and subsequently managed. One of the key considerations for building an effective Enterprise Metadata Repository requirement is that it allow for data quality assessment information – as well as data quality best practices, procedures, training materials, standards, and other data quality artifacts – to be readily available to future data quality projects, such that the information and experience from earlier efforts can be leveraged to yield greater success for each subsequent effort. Repository files should be available to the agency at-large, have the ability to be cross-referenced, and have some mechanism of version control.

The metadata repository supports the objectives of the FEA Data Reference Model by providing a consistent and reliable means of access to the principal artifacts of the DQI assessment efforts, including:

- Assessment reports,
- Data element inventory,
- Data quality worksheets with Content Quality and Data Management Maturity (Certification) ratings,
- Application inventory,
- Systems integration and data flows,
- Data models,
- Data dictionaries, and
- Business rule inventory.

The repository itself may be stored in a physical location or may be a virtual database, in which metadata is drawn from separate sources. Metadata may also include instructions for accessing specific data.

A database repository offers the solution with the most flexibility and options for retrieving data. Databases also offer more features that enable the design of customized security for a particular reviewer or groups of reviewers. An example of a model of a relational database to hold data quality assessments information is shown in Figure 4-6 below:

## Enterprise Metadata Repository – Data Model

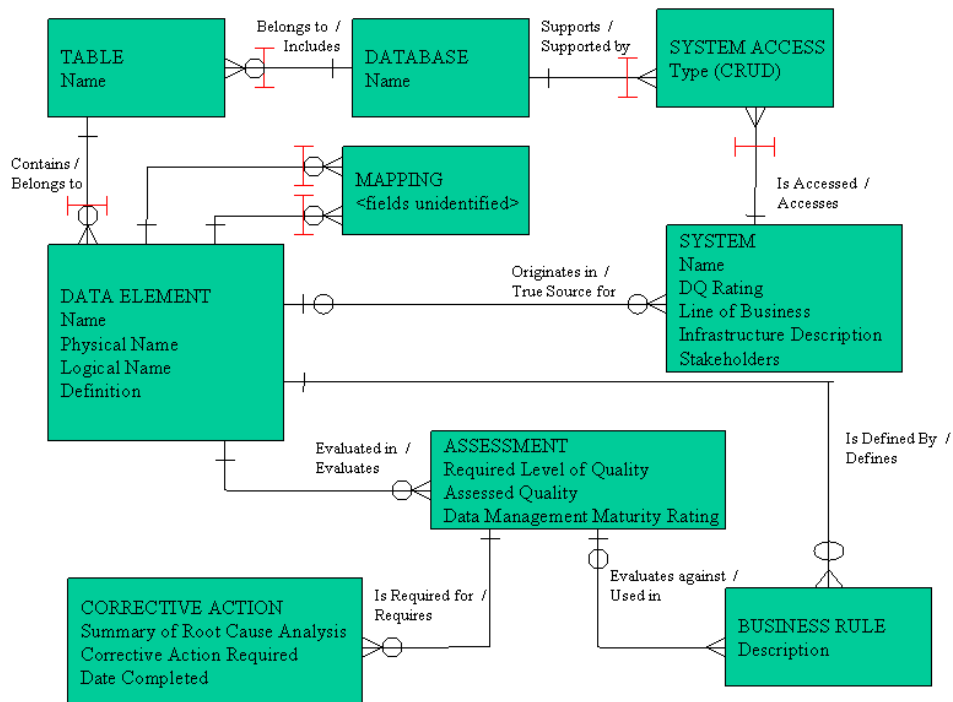


Figure 4-6: Enterprise Metadata Repository – Data Model

Once implemented and populated with assessment artifacts over time, a metadata repository provides the following value-added services to the enterprise:

- Enterprise Data Standards.** Having more than one set of standards diminishes, and frequently obliterates, the value of data assets across the enterprise. A Metadata Repository can provide standards and guidelines for all data objects, and can also yield examples of what are “good” and “bad” usage of the standards and guidelines. Standard data assets are:
  - Created once to avoid redundancy,
  - Defined consistently to ensure appropriate acquisition and use,
  - Documented consistently and stored to ensure the specifications can be easily found and used by those who require them, and
  - Used consistently, with minimal cost in all the areas where they are needed.
- Enterprise Glossary of Terms.** For an agency to achieve consistent, rigorous and actionable data definitions, it is imperative that all of the significant business concepts, common terms, common acronyms and common abbreviations be clearly and rigorously defined and easily accessible by all who are involved in the application development and maintenance processes, as well as those involved in the acquisition and use of the data. Where possible, these abbreviations should be based first on universally accepted, then on industry-wide, and last on enterprise-wide conventions.
- Enterprise Business Information Architecture (BIA):** This refers to an enterprise-wide data model that describes the data needs at the business level – it is equivalent to a Zachman Framework data model “Level 1” or “Level 2” (for secure areas). The BIA must be

- understood and approved by the agency's top management as the official representation of the major information subject areas or business resources. This model should be developed and maintained according to the agency's existing tool standard.
- **Enterprise Operational Data Model (ODM):** This refers to a model that describes the data needs at the business level but with additional detail for fundamental, highly shared enterprise data (master data). The ODM must be understood and approved by the agency's information stewards as the official representation of the fundamental information subject areas or business resources. It should also be developed using the standard tool.

## SECTION 5. DATA QUALITY TOOLS

Data quality tools provide automation and management support for solving data quality problems. The caveat for all automated correction tools is that some varying percentage of the data will need to be corrected and verified manually by looking at hard copy “official” documents or by comparing it to the real world object or event. Also, automated tools cannot ensure “completeness” or “accuracy” (see Glossary for definitions of these two dimensions).

Four major types of data quality tools and their usefulness are discussed below.

### **5.1 Data Profiling (Business Rule Discovery) Tools**

Data profiling tools may be used to analyze legacy system data files and databases in order to identify data relationships affecting the data. This analysis may identify quantitative (formula-based) or qualitative (relationship-based) conditions affecting the data and its successful migration and transformation. The analysis may also uncover exceptions or errors in the conditions. Data profiling previously was a difficult and tedious task requiring dozens of SQL programs searching through every record on every file or database to find data anomalies. Now this process is made easier through Business Rule Discovery tools.

For each column in a table, a data profiling tool will provide a frequency distribution of the different values, providing insight into the type and use of each column. Cross-column analysis can expose embedded value dependencies, while inter-table analysis explores overlapping values sets that may represent foreign key relationships between entities.

Data profiling can also be used to proactively test against a set of defined (or discovered) business rules. In this way, we can distinguish those records that conform to defined data quality expectations and those that do not, which in turn can contribute to baseline measurements and ongoing auditing for data quality reporting.

### **5.2 Data Defect Prevention Tools**

Automated tools may also be used to prevent data errors at the source of entry. Application routines can be developed that test the data input. Generalized defect prevention products enable the definition of business rules and their invocation from any application system that may use the data. These tools enforce data integrity rules at the source of entry, thereby preventing problems before they occur.

Proper use of data defect prevention tools begins by identifying the root causes for the data defects, which can be a combination of the following:

- Defective program logic,
- Not enough program edits,
- Not understanding the meaning of a data element,
- No common metadata,
- No domain definitions,
- No reconciliation process,
- No data verification process,
- Poor data entry training,
- Inadequate time for data entry, and
- No incentive for quality data entry.



The owners of the operational systems should plan to improve their programs and edit checks, unless the effort is unreasonably high. For example, if the corrective action requires changing the file structure, which means modifying (if not rewriting) most of the programs that access that file, then the cost for such an invasive corrective action on the operational system is probably not justified, especially if the bad data does not interfere with the operational needs of that system. This type of decision cannot—and should not—be made by IT alone. Downstream data consumers must negotiate with the data originators about justifying and prioritizing the data quality improvement steps.

### **5.3 Metadata Management & Quality Tools**

Metadata management refers to the activities associated with ensuring that metadata is properly created, stored, and controlled so that inconsistencies and redundancies can be removed. In short, metadata management is the act of imposing management discipline on the collection and control of metadata. The automated tools performing metadata management have the ability to:

- Capture metadata at the point of object creation,
- Store metadata in a common repository to allow for viewing and sharing resources or metadata across applications (“logical centralization”),
- Control inconsistencies and redundancies,
- Ensure conformance to data naming standards (such as the ISO/IEC 11179 Metadata Registry standard),
- Maintain metadata for control of data reengineering and correction processes,
- Evaluate data models for normalization, and
- Evaluate database design for integrity, such as primary key to foreign key integrity, and performance optimization.

While many tools performing metadata management are essentially repositories (CRM tools, “incident tracking” tools, etc.), a new class of powerful tools uses a real-time architecture to provide data managers with reliable information to make near-real-time decisions and to give developers details about web services so they can reuse them in service-oriented architectures.

Metadata management and quality tools support the documentation of the specification of information products. These tools cannot determine if data required for proper job performance and execution is missing, defined correctly, or even required in the first place. Information resource data (metadata) quality tools may audit or ensure that data names and abbreviations conform to standards, but they cannot assess whether the data standards are “good” standards that produce data names that are understandable to the enterprise.

### **5.4 Data Reengineering and Correction Tools**

Data reengineering and correction tools may be used either to actually correct the data or to flag erroneous data for subsequent correction. These tools require varying degrees of in-house data knowledge and analysis to adequately use them. Data correction tools may be used to standardize data, identify data duplication, and transform data into a correct set of values. These tools are invaluable in automating the most tedious facets of data correction.

Data reengineering and correction tools may perform one or more of the following functions:

- Extracting data.
- Standardizing data.
- Matching and consolidating duplicate data.
- Reengineering data into architected data structures.
- Filling in missing data, based upon algorithms or data matching.

- Applying updated data, such as address corrections from change of address notifications.
- Transforming data values from one domain set to another.
- Transforming data from one data type to another.
- Calculating derived and summary data.
- Enhancing data, by matching and integrating data from external sources.
- Loading data into a target data architecture.

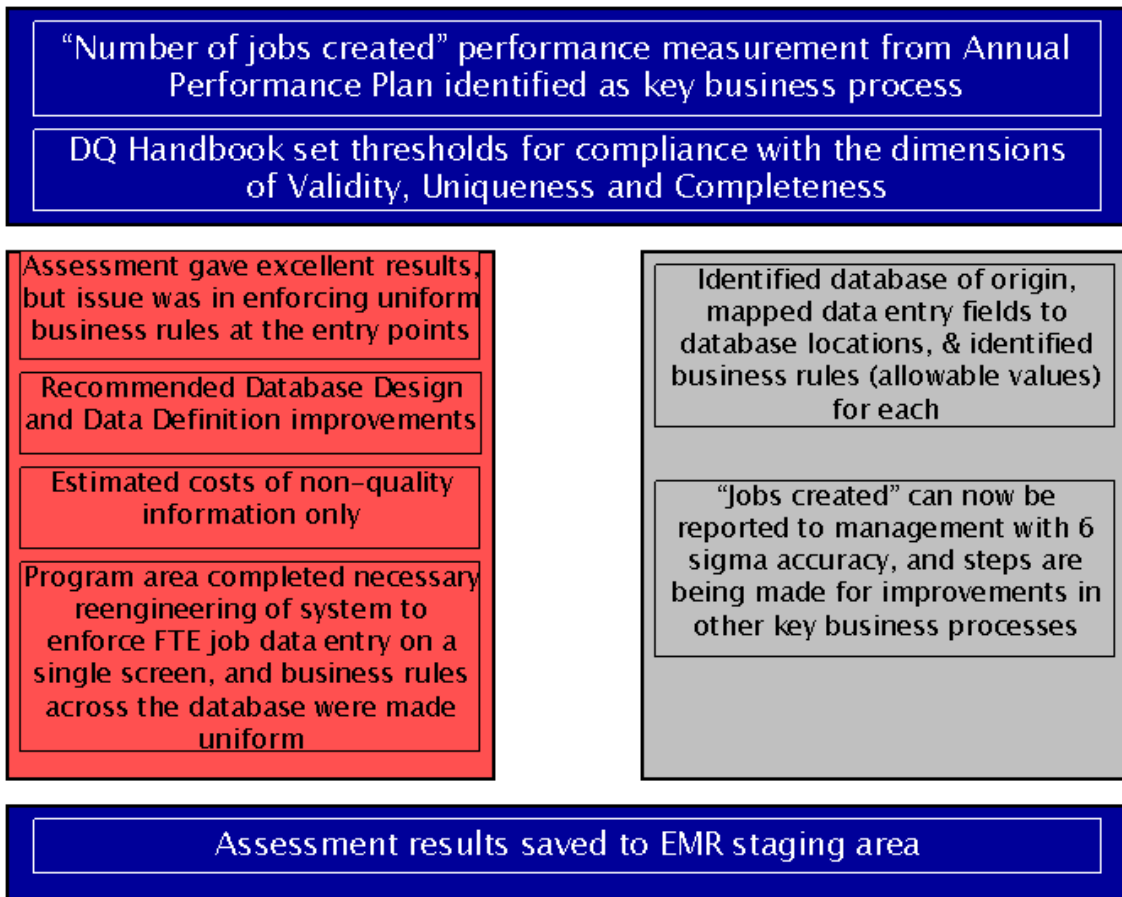
**APPENDIX A. EXAMPLES OF DQI AT FEDERAL AGENCIES**

The purpose of this Appendix is to provide two federal agency examples of DQI in current practice. The DQI Process Framework (see Figure 4-1) is used to measure the extent and level of DQI being practiced, and a scorecard is provided for each agency marking their DQI successes as well as continued challenges in pursuing their respective data quality objectives.

**A.1 Department of Housing and Urban Development**


At the Department of Housing and Urban Development (HUD), data quality initiatives focused on a major legacy information system charged with measuring job creation in underprivileged areas. Data to support this measurement were collected from three different points in the web-based application. During the assessment of the database where the three classes of job-creation data were stored, it was determined that the three data entry points did not use a uniform method of collection, so that only a portion of the data could be converted to full-time equivalent (FTE) jobs created. It was further determined that an information architecture redesign was required to better support the accuracy and quality of the information exchange between on-line application and database.

Figure A.1 below shows the DQI processes employed in this data quality assessment effort:



**Figure A.1: HUD DQI Implementation**

Following the assessment, a DQI scorecard was prepared to mark the successes of the effort and highlight the areas where further progress was required (Figure A.2):

	<b>Enterprise Level</b> (some DQI impact felt here)	<b>Program Level</b> (modest DQI impact felt here)	<b>System Level</b> (most DQI impact here)
<b>Successes</b>	1. Annual Performance Plan effective blueprint for identifying key business processes/ data sources 2. Development of DQ Handbook with consistent standards and DQI procedures 3. Data Control Board created for DQ governance	1. Reengineered system to 6 sigma for this metric 2. Information Value Cost Chain completed for in-scope data showing transformations, data classes, and system interfaces	1. Costs of non- quality information estimated 2. Information Architecture alignment with database improved 3. System functionality improved 4. New Data Dictionary developed
<b>Challenges remaining</b>	1. EDM staging area not secure, robust enterprise solution required 2. Training required across the enterprise	1. Data Quality Assurance plan not formalized 2. Root Cause Analysis not undertaken – errors may return and impact other business processes 3. DQ stewardship lacking at program level	1. Lack of Statistical Process Control 2. Database partitioned between grants programs, resulting in data overlap and lack of visibility

**Figure A.2: HUD DQI Scorecard**

**A.2 Defense Logistics Agency**

At the Defense Logistics Agency (DLA), data quality initiatives focused on building an understanding of data and functional process flows of four feeder data systems into a DLA portal under development. Multiple data entry points of the same classes of mission-critical data were analyzed to determine Authoritative Data Sources. In addition, an investigation into data stewardship responsibilities was undertaken.

Figure A.3 below shows the DQI processes employed in this data quality assessment effort:

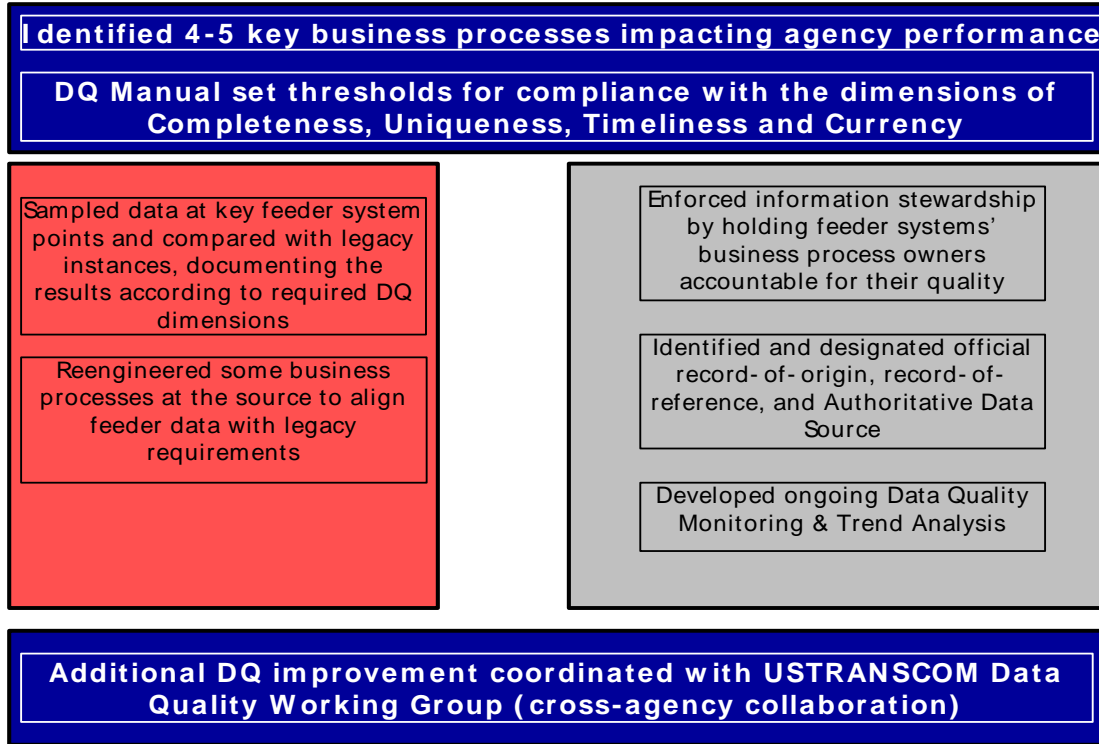


Figure A.3: DLA DQI Implementation

Following the assessment, a DQI scorecard was prepared to mark the successes of the effort and highlight the areas where further progress was required (Figure A.4):


	<b>Enterprise Level</b> (minimal DQI impact felt here)	<b>Program Level</b> (most DQI impact felt here)	<b>System Level</b> (modest DQI impact here)
<b>Successes</b>	1. Some key business processes and their sequencing (operational "racetrack") developed for first time 2. DQ Manual developed with metrics and standards	1. Data Integrity Branch (DIB), program area stewardship defined 2. Data Quality Monitoring & Trend Analysis program taken up by DIB	1. Assessment points for sampling feeder data developed strategically 2. Reengineered some business processes to decrease data redundancy
<b>Challenges remaining</b>	1. EMD Repository solution required 2. Training required across the enterprise 3. Future DQ issues to be identified/ coordinated with USTRANSCOM Data Quality Working Group (cross-agency collaboration)	1. Authoritative Data Source (ADS) analysis completed, but full information Value Cost Chain from feeders to legacy not understood	1. Refining Statistical Process Control methodology 2. Determining ROI for DQ improvement 3. Defining investment threshold for reaching point of diminishing return

Figure A.4: DLA DQI Scorecard

## APPENDIX B. EVOLUTION OF INFORMATION QUALITY MANAGEMENT

The United States manufacturing industry operated in a steady state from the end of World War II until the late 1970's, when it suffered a revolution caused by a redefinition of quality. The new paradigm of quality owed its creation to the Japanese manufacturing industry's application of Dr. W. E. Deming's principles of quality. Before this revolution, quality was thought to be "product intrinsic" and therefore achievable by after-the-fact inspection (the "quality control" school of thought). If the product was defective, it was either sent back for correction (re-worked) or disposed of (scrapped).

However, this approach directly increased costs in three ways: first, the added cost of inspection; second, the cost of re-work; third, the cost of scrap. In those cases where inspection was based on samples (not 100% inspections), there were also the costs of delivering a defective product to a customer (including dissatisfaction and handling of returns). Dr. Deming questioned the quality control approach and affirmed that quality can best be achieved by designing it into a product and not by inspecting defects out of a finished product. He indicated that inspection should be used at a minimum and only to determine if there is unacceptable variability, and advocated a focus on improving the process in order to improve the product. Also, he centered his definition of quality on the customer, not the product. He indicated that quality is best measured by how well the product meets the needs of the customer.

Dr. Deming's approach, used since the early 1960's,<sup>5</sup> was also based on the "PDCA" approach (continuous process improvement) developed by W. Shewhart,<sup>6</sup> and the Total Quality Management approach developed by P. B. Crosby.<sup>7</sup> M. Imai incorporated the proactive PDCA approach in his Kaizen and Gemba Kaizen methods of continuous process improvement in which everyone in the organization is encouraged to improve value-adding processes constantly to eliminate the waste of scrap and rework, and in which improvements do not have to cost a lot of money.<sup>8</sup>

---

<sup>5</sup> W. E. Deming, *Out of the Crisis*: Cambridge: MIT Center for Advanced Engineering Study, 1986.

<sup>6</sup> W. Shewhart, *Statistical Method from the Viewpoint of Quality Control*: New York: Dover Publications, 1986.

<sup>7</sup> P. B. Crosby, *Quality is Free: The Art of Making Quality Certain*: New York: Penguin Group, 1979.

<sup>8</sup> M. Imai, "Kaizen: The Key to Japan's Competitive Success:" New York, Random House, 1989; and "Gemba Kaizen: Low Cost Approach to Management:" New York: McGraw-Hill, 1997.

## APPENDIX C. GLOSSARY

**6-sigma (6 $\sigma$  or 6s):** Six standard deviations used to describe a level of quality in which six standard deviations of the population fall within the upper and lower control limits of quality with a shift of the process mean of 1.5 standard deviations, and in which the defect rate approaches zero, allowing no more than 3.4 defects per million parts.

**Accessibility:** the degree to which internal or external customers are able to access or get the information they need.

**Accuracy to reality:** A characteristic of data quality measuring the degree to which a data value (or set of data values) correctly represents the attributes of the real-world object or event.

**Atomic level:** Defines attributes that contain a single fact. For instance, “Full Name” is not an atomic level attribute because it can be split into at least two distinct pieces of information: “First Name” and “Last Name.”

**Authoritative data source (ADS):** A cohesive set of data assets that provide trusted, timely and secure information to support a business process.

**Business concept:** A person, place, thing, event or idea that is relevant to the business and for which the enterprise collects, stores, and applies information.

**Business rule:** A statement expressing a policy or condition that governs business actions and establishes data integrity guidelines.

**Community of Interest (COI):** Consists of collaborative groups who come together to solve information sharing problems in order to develop operational capabilities. COI solve information sharing problems by developing a shared vocabulary to exchange information in pursuit of their shared goals, interests, missions, or business processes.

**Completeness:** A characteristic of data quality measuring the degree to which all required data are known.

**Concurrency:** A characteristic of data quality measuring the degree to which the timing of equivalence of data is stored in redundant or distributed database files. Data concurrency may describe the minimum, maximum, and average information float time from when data are available in one data source and when they become available in another data source

**Consistency:** A measure of data quality expressed as the degree to which a set of data is equivalent in redundant or distributed databases.

**Contextual clarity:** the degree to which information presentation enables internal or external customers to understand the meaning of the data and avoid misinterpretation.

**Data:** The representation of facts. Data can represent facts in many mediums or forms including digital, textual, numerical, or graphical. (1) *Raw data* are data units that are used as the raw material in a predefined process that will ultimately produce an information product (e.g. single number, file, report, image, verbal phrase). (2) *Component data* are a set of temporary, semi-processed information needed to manufacture the information product (e.g. file extract, intermediary report, semi-processed data set).

**Data administrator:** The individuals who are responsible for establishing the linkages between source data, information products and policies or regulations that enforce how data and information can be used, and for establishing information policy. Often the senior individuals at the program office level, Data Administrators have ultimate responsibility for ensuring accuracy, completeness, validity and reproducibility of data stored in systems used to support the program office lines of business.

**Data definition:** The process of analyzing, documenting, reviewing, and approving unique names, definitions, characteristics and representations of data according to established procedures and conventions and standards.

**Data dictionary:** A repository of information (metadata) defining and describing the data resource. An *active* data dictionary, such as a catalog, is one that is capable of interacting with and controlling the environment about which it stores information or metadata. An *integrated* data dictionary is one that is capable of controlling the

data and process environments. A *passive* data dictionary is one that is capable of storing metadata or data about the data resource, but is not capable of interacting with or controlling the computerized environment external to the data dictionary.

**Data element:** The smallest unit of named data that has meaning. A data element is the implementation of an attribute.

**Data profiling:** The use of analytical techniques about data for the purpose of developing a thorough knowledge of their content, structure and quality.

**Data quality:** The state of excellence that exists when data are relevant to their intended uses, and are of sufficient detail and quantity, with a high degree of accuracy and completeness, consistent with other sources, and presented in appropriate ways.

**Data quality assessment:** The random sampling of a collection of data and testing it against its valid data values to determine its accuracy and reliability.

**Data standardization:** The process of achieving agreement on common data definitions, representation, and structures to which all data layers must conform.

**Data steward:** Business or technical persons or group that manages the development, approval, and use of data within a specified functional area, ensuring that it can be used to satisfy business data requirements throughout the organization. Data Stewards have ultimate responsibility for defining data requirements.

**Database of Origin:** The first electronic file in which an occurrence of an entity type is created and stored. Also known as *Record of Origin*.

**Database of Record:** The single, authoritative database file for a collection of fields for occurrences of an entity type. This file represents the most reliable source of operational data for these attributes or fields. Also known as *Record of Reference*.

**Defect:** An item that does not conform to its quality standard or customer expectation.

**Derivation integrity:** the correctness with which derived data are calculated from their base data.

**Derived data:** Data that is created or calculated from other data within the database or system.

**Domain:** (1) Set or range of valid values for a given attribute or field, or the specification of business rules for determining the valid values. (2) The area or field of reference of an application or problem set.

**Federal Enterprise Architecture (FEA):** An initiative of the Office of Management and Budget (OMB) that aims to comply with the Clinger-Cohen Act and provide a common methodology for information technology acquisition and for describing information technology resources in the United States Federal government. Includes the Performance Reference Model (PRM), the Business Reference Model (BRM), the Service Component Reference Model (SRM), the Data Reference Model (DRM) and the Technical Reference Model (TRM).

**Format consistency:** The use of a standard format for storage of a data element that has several format options. For example, Social Security Number may be stored as the numeric “123456789” or as the character “123-45-6789”. The use of a uniform format facilitates the comparison of data across databases.

**Information:** (In the context of information dissemination by federal agencies, appropriate for the DAS DQ Framework). Any communication or representation of knowledge such as facts or data, in any medium or form, including textual, numerical, graphic, cartographic, narrative, or audiovisual forms. This definition includes information that an agency *disseminates* from a web page, but does not include the provision of hyperlinks to information that others disseminate. This definition does not include opinions, where the agency's presentation makes it clear that what is being offered is someone's opinion rather than fact or the agency's views.

**Information architecture:** A “blueprint” of an enterprise expressed in terms of a business process model, showing what the enterprise does; an enterprise information model, showing what



information resources are required; and a business information model, showing the relationships of the processes and information.

**Information float:** The length of the delay in the time a fact becomes known in an organization to the time in which an interested internal customer is able to know that fact. Information float has two components: Manual float is the length of the delay in the time a fact becomes known to when it is first captured electronically in a potentially sharable database. Electronic float is the length in time from when a fact is captured in its electronic form in a potentially sharable database, to the time it is “moved” to a database that makes it accessible to an interested customer.

**Information group:** A relatively small and cohesive collection of information, consisting of 3–25 data elements and related entity types, grouped around a single subject or subset of a major subject. An information group will generally have one or more subject matter experts and several business roles that use the information.

**Information producer:** The role of individuals in which they originate, capture, create, or maintain data or knowledge as a part of their job function or as part of the process they perform. Information producers create the actual information content and are accountable for its accuracy and completeness to meet all information stakeholders’ needs.

**Information quality:** The quality of the content of information systems, ensuring that the data presented has value and models the real world.

**Information stakeholder:** Any individual who has an interest in and dependence on a set of data or information. Stakeholders may include information producers, knowledge workers, external customers, regulatory bodies, and various information systems roles such as database designers, application developers, and maintenance personnel.

**Information value/cost chain:** The end-to-end set, beginning with suppliers and ending with customers, of processes and data stores, electronic and otherwise, involved in creating, updating, interfacing, and propagating data of a type from its origination to its ultimate data store, including independent data entry processes, if any.

**Integrity:** The security of information; protection of the information from unauthorized access or revision, to ensure that the information is not compromised through corruption or falsification.

**ISO/IEC 11179:** The international standard for representing metadata for an organization in a Metadata Registry.

**Line of sight:** The indirect or direct cause and effect relationship from a specific IT investment to the processes it supports, and by extension the customers it serves and the mission-related outcomes to which it contributes.

**Metadata:** A term used to mean data that describes or specifies other data. The term *metadata* is used to define all of the characteristics that need to be known about data in order to build databases and applications and to support internal/external customers and information producers.

**Non-duplication:** A characteristic of data quality measuring the degree to which there are no redundant occurrences of data.

**Precision:** the degree to which data are known to the right level of granularity (e.g., the right number of decimal digits right of the decimal point, time to the hour or the half-hour or the minute, or the square footage of a building is known to within one square foot as opposed to the nearest 100s of feet).

**Real-time:** Pertaining to the timeliness of data or information which has been delayed only by the time required for electronic communication. This implies that there are no noticeable delays.

**Record of origin:** The first electronic file in which an occurrence of an entity type is created.

**Record of reference:** The single, authoritative database file for a collection of fields for occurrences of an entity type. This file represents the most reliable source of operational data for these attributes or fields. In a fragmented data environment, a single occurrence may have different collections of fields whose record of reference is in different files.

**Referential integrity:** Integrity constraints that govern the relationship of an occurrence of one entity type or file to one or more occurrences of another entity type or file, such as the relationship

of a customer to the orders that customer may place. Referential integrity defines constraints for creating, updating, or deleting occurrences of either or both files.

**Relationship validity:** The degree to which related data conforms to the associative business rules.

**Repository:** A database for storing information about objects of interest to the enterprise, especially those required in all phases of database and application development. A repository can contain all objects related to the building of systems including code, objects, pictures, definitions. The repository acts as a basis for documentation and code generation specifications that will be used further in the systems development life cycle.

**Rightness or fact completeness:** The degree to which the information presented is the right kind and has the right quality to support a given process or decision.

**Shewart Cycle (PDCA):** PDCA ("Plan-Do-Check-Act") is an iterative four-step problem-solving process typically used in quality control. Plan: establish the objectives and processes necessary to deliver results in accordance with the specifications. Do: implement the processes. Check: monitor and evaluate the processes and results against objectives and Specifications and report the outcome. Act: apply actions to the outcome for necessary improvement.

**Service Oriented Architecture (SOA):** a computer systems architectural style for creating and using business processes, packaged as services, throughout their lifecycle. SOA separates functions into distinct units (services), which can be distributed over a network, thereby combined and reused to create business applications.

**Timeliness:** A characteristic of data quality measuring the degree to which data are available when internal/external customers or processes require it.

**UML (Unified Modeling Language):** A standard notation for the modeling of real-world objects as a first step in developing an object-oriented design

methodology. Among the concepts of modeling that UML specifies are class (of objects), object, association, responsibility, activity, interface, use case, package, sequence, collaboration, and state.

**Usability:** the degree to which the information presentation is directly and efficiently applicable for its purpose.

**Utility:** The usefulness of the information to its intended consumers, including the public. In assessing the usefulness of information that the agency disseminates to the public, the agency needs to consider the uses of the information not only from the perspective of the agency but also from the perspective of the public. As a result, when transparency of information is relevant for assessing the information's usefulness from the public's perspective, the agency must take care to ensure that transparency has been addressed in its review of the information.

**Validity:** A characteristic of information quality measuring the degree to which the data conforms to defined business rules. Validity is not synonymous with *accuracy*, which means the values are the correct values. A value may be a valid value, but still be incorrect. For example, a customer date of first service can be a *valid* date (within the correct range) and yet not be an *accurate* date.

**APPENDIX D. ADDITIONAL REFERENCES**

- Thomas C. Redman, *Data Quality for the Information Age* (Norwood, MA: Artech House Inc., 1996).
- Thomas C. Redman, *Data Quality – The Field Guide* (Digital Press, 2001).
- Larry P. English, *Improving Data Warehouse and Business Information Quality (IDW&BIQ)* (New York: John Wiley & Sons, 1999).
- Richard Y. Wang, Mostapha Ziad, Y.W. Lee, *Data Quality* (Springer, 2000).
- Yang W. Lee, Leo L. Pipino, James D. Funk, Richard Y. Wang, *Journey to Data Quality* (MIT Press, 2006).
- W. Edwards Deming, *Quality, Productivity and Competitive Position* (Cambridge: MIT CAES, 1982).
- J.M. Juran, *Managerial Breakthrough* (New York: McGraw-Hill, 1964).
- Walter A. Shewart, *Statistical Method from the Viewpoint of Quality Control* (Mineola: Dover Publications, 1986).
- David Loshin, *Enterprise Knowledge Management - The Data Quality Approach* (San Francisco: Morgan Kaufmann, 2001).
- Ralph Kimball and Joe Caserta, *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data* (Wiley, John & Sons, Incorporated, 2004).
- Jack Olson, *Data Quality: The Accuracy Dimension* (San Francisco: Morgan Kaufman Publishers, 2001).
- Nancy R. Tague, *The Quality Toolbox, 2<sup>nd</sup> Edition* (American Society for Quality, 1995).