United States Department of the Interior

Data Quality Management Guide

August 2008

Office of the Chief Information Officer

FOREWORD

The purpose of the Department of the Interior's (DOI) Data Quality Management Guide is to provide a repeatable set of processes for monitoring and correcting the quality of data in DOI-owned data sources, in particular Authoritative Data Sources.

This Guide is a companion to the OMB Information Quality Guidelines pursuant to Section 515 of the Treasury and General Government Appropriations Act for Fiscal Year 2001 (http://www.doi.gov/ocio/guidelines/515Guides.pdf). The Guide addresses in detail the processes required to ensure measurable quality in DOI-owned source data, helping to ensure that quality source data are available for use in DOI's information dissemination products. The quality of the disseminated information products is managed by DOI's Information Quality Guidelines, established in accordance with the OMB Information Quality Guidelines and cited above.

Section 515 of the Treasury and General Government Appropriations Act for Fiscal Year 2001 is commonly known as the Information Quality Act (IQA). The IQA requires federal agencies to develop procedures for ensuring the quality, objectivity, utility, and integrity of the information they disseminate. Included are administrative mechanisms allowing affected persons to seek and obtain correction of information maintained and disseminated by federal agencies. The agencies report to OMB annually on complaints they received under the IQA and the disposition of those complaints.

DOI's Data Quality Management Guide provides tangible business benefits for the DOI because it:

- Relies on government- and industry-accepted practices of applying data quality criteria to ensure a consistent level of quality.
- Facilitates Department-wide monitoring of correction and improvement activities that are not related to data correction requests made under the IQA.
- Facilitates integration and coordination of data quality activities with other data management activities.
- Provides common milestones and products for data quality management activities.
- Enables the institutionalization of data quality management.
- Provides the capability to share techniques, solutions, and resources throughout the Department.

The intended audience for this Guide is largely Information Technology (IT) practitioners who are responsible for monitoring and correcting the quality of data in data sources owned and managed by DOI. IT practitioners coordinate with principal data stewards and business data stewards to understand their expectations for the desirable level of quality and scope of data to be measured.

Send recommended changes to the document to:

Senior Information Architect
E-mail: iea@ios.doi.gov

Table of Contents

Table of Figures

Table of Tables

INTRODUCTION

DOI's Data Resource Management and Standardization Program was established pursuant to Data Quality legislation as part of the FY 2001 Consolidation Appropriations Act (Public Law 106-554 section 515). Building upon the Data Quality report language contained in the FY 1999 Omnibus Appropriations Act (Public Law 105-277[1]), the Act directed the Office of Management and Budget (OMB) to issue government-wide guidelines that "provide policy and procedural guidance to federal agencies for ensuring and maximizing the quality, utility, and integrity of information (including statistical information) disseminated by federal agencies."

The statute further requires that each agency issue its own guidelines for implementation of PL 106-554, section 515. Further, it requires that an administrative process be established to allow affected parties to seek correction of agency information that does not meet or comply with OMB guidelines. OMB responded with a notice to all federal agencies (67 Federal Regulation 369, January 3, 2002) requiring that they have their own information guidelines in place by October 1, 2002. The agency guidelines must apply the statutory requirements cited above to their own particular programs. To do so, each agency must adopt a basic standard of data quality that is specific to the data's usage and supporting information quality.

DOI's Data Resource Management and Standardization Program is responsible for developing Department-wide data architecture, data governance and stewardship infrastructure, and promoting data quality improvement. This document will focus on the Data Quality Improvement Process prior to data dissemination in support of the OMB Information Quality Guidelines.

The DOI Data Quality Improvement Process follows a four-staged process: assessment, process improvement, correction, and certification. The System Owners, Principal Data Stewards, and Bureau Business Data Stewards coordinate their efforts for performing their own assessment, process improvement, data correction, and certification. DOI's Senior Information Architect provides guidance, as needed.

The primary objective of the Data Quality Improvement Process is to improve the quality of **mission-critical** data in the major information systems that DOI owns and manages. Mission-critical data are data that are essential for DOI to conduct its business, data frequently used by the Department, data that are key to the Department's integrity and accountability, and data used to support Government Performance and Results Act (GPRA) reports.

This Data Quality Management Guide provides a description of the processes needed to guide the efforts of DOI's organizations for continuous data quality improvement in the acquisition, creation, maintenance, storage, and application of data. This Guide provides general technical guidance for all data communities at DOI (e.g., geospatial, law enforcement, finance, recreation, and facility management). The Guide is written primarily for an IT audience that would be conducting a data quality assessment directed and coordinated by the business area leaders responsible for improvement in data quality. This Guide prescribes standard processes that can be adapted and extended for data communities, particularly geospatial and other scientific data communities, with specialized data quality management requirements. It complements DOI's Data Resource Management[2] policy that establishes the need for DOI's organizations to seek alignment of data management priorities with DOI's mission and data quality project objectives. It also complements the requirements outlined in PL 106-554, section 515, as it provides a process to improve the quality of data used in information prior to the dissemination of this information to the public.

This Guide contains the concepts, step-by-step processes, and an illustration of and references to worksheets that will facilitate data quality improvement and correction processes. The Sections of this Guide are:

- **Chapter 1: Implementing DOI's Data Quality Improvement Environment**. Provides the definition of data, the quality standard, and roles and responsibilities; distinguishes data quality from information quality; and gives an overview of the Data Quality Improvement method.
- **Chapter 2: Data Quality Assessment Process**. Describes the assessment process from selection to final report.
- **Chapter 3: Data Quality Improvement Process**. Describes the methods and techniques that should be used by the data quality project personnel to perform Data Quality Improvement.
- **Chapter 4: Data Quality Correction Process**. Describes the methods and techniques that should be used by the data quality project personnel to perform data correction.
- **Chapter 5: Data Quality Certification Process**. Describes the methods and techniques that will be used to perform the final task of independent verification or certification of mission-critical information.
- **Appendix A: Data Quality Improvement Process Planning**. Contains guidance and examples of work breakdown structures to plan data quality assessments, data quality improvements, data corrections, or certification projects.
- **Appendix B: Data Quality Software Tools**. Contains a discussion of available software tools that can be used to automate data quality management and improvement.
- **Appendix C: Data Quality Improvement Process Background**. Provides a brief background description on the evolution of the critical concepts related to Total Quality Management and the Data Quality Improvement Process in general.
- **Appendix D: The "Acceptable Quality Level" Paradigm**. Provides an explanation of how to determine an acceptable level of data quality.
- **Appendix E**: **Additional Legislation/Regulations Influencing DOI's Guide.** Provides a list of additional legislation and/or regulations that support DOI's Data Quality Management.
- **Appendix F: Glossary.** Provides a glossary of terms used in this document that require special attention.
- **Appendix G: Endnotes.**

Chapter 1. IMPLEMENTING DOI'S DATA QUALITY IMPROVEMENT PROCESS ENVIRONMENT

1.1 Overview

This chapter presents a definition of data and data quality, DOI's data quality standard, roles and responsibilities, and an overview of the data quality improvement process method. The implementation of this method must be in the context of business, organizational, and cultural transformation characterized by:

- A set of goals: "Focus resources on DOI's/Bureaus' data quality objectives."
- A value system: "We value our data customers."
- A mindset: "Data are key sharable assets and are used in important information products for citizens."
- An environment that promotes continuous improvement: "To eliminate the waste associated with process failures and rework caused by defective data."

1.2 Definition of Data

According to the Federal Enterprise Architecture Data Reference Model[3], a datum (pl. 'data') is a value, or set of values, representing a specific concept or concepts. Data become 'information' when analyzed and possibly combined with other data in order to extract meaning and to provide context. The representation of data may be captured in many media or forms including written, digital, textual, numerical, graphical, audio, and video. DOI's Data Quality Management Guide provides a method for implementing data quality management applicable to all forms of data required by DOI.

1.3 Definition of Data Quality

The definition given for data quality by J.M. Juran[4], author of *Juran's Quality Handbook*, is the one in widest use today: "Data are of high quality if they are fit for their intended uses in operations, decision making and planning." Data quality means that 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 is not the same thing as information quality. While *data quality* ensures that data are fit for their intended uses in operations, decision making and planning, *information quality* describes the quality of the content of information systems, ensuring that the data presented have value and model the real world correctly for its planned use. Information is the meaning given to data or the interpretation of data based on its context; it is the finished product that results from this interpretation. In Section 515 guidelines, the term information is used primarily in the context of dissemination of information and correction of disseminated information.

Within DOI, data are considered the most fundamental units of "manufactured" information that are subsequently assembled by one or multiple processes (e.g., request for park permit or an arrest in a park) and consumed by other processes (e.g., reporting performance indicators) or customers (e.g., park resource scheduling coordinator). Customers consume the information product (i.e., data purposefully assembled) to make critical decisions, such as underwriting an application or securing funding for future programs, providing insight into the Department's performance, or servicing a public need.

Improving data quality involves correcting defective data and implementing quality improvement procedures that ensure that the expected levels of data quality are achieved and maintained. The two principal subsets of data quality that will be measured and reported using the instructions articulated in this Guide are the following: (See Table 1-2 for measurement standards.)

- **Data Definition and Data Architecture Quality**. Proper data definition accurately describes the meaning of the real-world object or event that the data represent and meets the needs of the information customers to understand the data they use. Proper data architecture correctly represents the structure of the inherent and real relationships of data to represent real-world objects and events and is stable and flexible. Data definition and data architecture are the specification of the information product and must represent the views and needs of the business areas, applications, and end users of the information. Data definition and architecture include the business definition, the domain or value set, and the business rules that govern the data. For detailed descriptions of data definition and data architecture dimensions, refer to Section 2.2.
- **Data Content Quality**. Data content quality cannot be measured without a quality definition. Data content quality is the degree to which the data values accurately represent the dimensions of the real-world object or event and meet the needs of the information end users to perform their jobs effectively.

1.4 OMB's Information Quality Guidance and DOI's Data Quality Guidelines

Based on guidance provided in OMB Section 515, PL 106-554, agencies are directed to develop management procedures for reviewing and substantiating the quality of data before they are disseminated. OMB directed agencies disseminating influential scientific, financial, or statistical information to ensure that the original or supporting data are generated and that the analytical results are developed using sound statistical methods. In OMB Section 515 guidelines, the term data is used primarily in the context of dissemination of data and correction of disseminated data. Given that the method described in this Guide applies to data before they are disseminated, it can be applied to the improvement and correction of data as needed.

Data quality processes are different from those for information quality. Organizations must establish standards and guidelines for data managers, data stewards, and anyone with an interest in the data to ensure that data quality is addressed during the entire lifecycle of data's movement through information systems. 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, resolution procedures for data disputes, roles and responsibilities, and standard documentation formats. All of these policies, procedures, and definitions are part of the definition of data quality in the context of DOI's Data Quality Management Guide. The quality measures in this Guide complement OMB's Information Quality Guidelines as follows:

OMB's Information Quality Dimensions	**OMB Definition Quoted in Part from OMB's Guidelines (http://www.whitehouse.gov/omb/fereg/reproducible2.pdf)**	**DOI's Data Quality Guide's Granular Measures Supporting OMB's Guidelines (For definitions, see page 2-4.)**
Utility	Refers to the usefulness of the information to its intended users, including the public	Timeliness, Concurrency, Precision

Objectivity	Involves two distinct elements, i.e., presentation and substance	
	(a) Includes whether disseminated information is being presented in an accurate, clear, complete, and unbiased manner	Information Quality measures not addressed in DOI's Guide
	(b) Involves a focus on ensuring accurate, reliable, and unbiased information	Accuracy to Reality, Accuracy to Original Source, Precision, Validity, Completeness, Relationship Validity, Consistency, Concurrency and Derivation Integrity
Integrity	Refers to the security of information, i.e., 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 DOI's Guide

Table 1-1: DOI's Data Quality Dimensions Mapped to OMB's Information Quality Dimensions

1.5 DOI's Bureaus' Mission-Critical Data Quality Standards

Each Bureau must define its data quality standards based upon internal users' and external stakeholders' expectations and requirements for the data. In order to provide the Data Quality Project Team (i.e., an ad-hoc, federated Bureau team consisting of the Authoritative Data Source (ADS) steward and a group of data experts charged with assessing, correcting, and certifying data) with specific and actionable direction, this section describes specific standards for mission-critical data. This framework should be used to establish data quality standards for other data deemed critical by the Data Quality Project Team.

The two areas of measurable data quality must be managed based on the quality class of the data to achieve total data quality. 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 sigma level corresponding to each data quality class is indicated below and defined in the Glossary of this Guide. The three data quality classes are:

- **Absolute Tier (Near Zero-Defect Data).** Indicates these data can cause significant process failure if they contain defects. Institutionalizing this tier into the fabric of DOI's data architecture would require significant investments in training and infrastructure. Therefore, the standard should only be applied to data so critical that the consequences of incorrectness could mean significant losses.
- **Second Tier (High-Quality Data).** 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 (Medium-Quality Data).** Indicates the costs associated with defects in these data are moderate and should be avoided whenever possible. Quality tolerance for this level should be no worse than 3 sigma, or 66,807 errors per million (a 6.7% error rate).

The Data Quality Project Team determines the scope of the data to be reviewed during the Data Quality Assessment phase described in this Guide. When there is no impact associated with data defects, it may be an indication that the Department does not require the data at all.

Table 1-2 shows DOI's quality standards for **mission-critical** data, which is class Second Tier.

Measure		Target	Confidence Level
Definition	% Complete	100%	99%
	Customer Satisfaction	100%	99%
	Known and Acceptable Definition and Structure Defects	100%	99%
Content	Data Correctness	4σ	99%
	Processes Producing Mission Critical Information in Statistical Control	100%	99%
	Elimination of Known Defect Production (New Defects) Through Information Quality Improvement	50% (per year)	N/A

Table 1-2: DOI's Mission-Critical Data Quality Standards

The data content quality standard in Table 1-2 applies to controllable, mission-critical data. Controllable means that each Bureau has control over the content, because it is collected following Bureau standards (e.g., recreation reservation) or produced within the Bureau. The short- and long-term targets for data content quality error rates assume the commonly accepted allowance that a process mean could shift by 1.5 standard deviations.

1.6 Data Quality Improvement Process Roles and Responsibilities

This Guide will not define the Data Quality Improvement Process roles and responsibilities, since these will be determined by the Bureau Data Architect, Principal Data Stewards, Business Data Stewards, and the Bureaus' Data Quality Project Teams. These teams should include system owners, subject matter experts, and data architects who understand the intended use of the data across the business processes that take action on the data from initial creation to final disposal. Overall DOI's data governance roles and responsibilities are described in the document, *DOI Data Standardization Procedures*[5] (April 2006). Section 2.2.5.9 of that document defines the Bureaus' Business Data Stewards' data quality responsibilities. Figure 1-1 illustrates DOI's data governance roles and relationships, which include roles such as DOI's Data Architect, DOI's Data Advisory Committee, Executive Sponsor, Principal Data Steward, Bureau Business Data Steward, Subject Matter Experts, Bureau Data Architect, and Bureau Database Administrator.

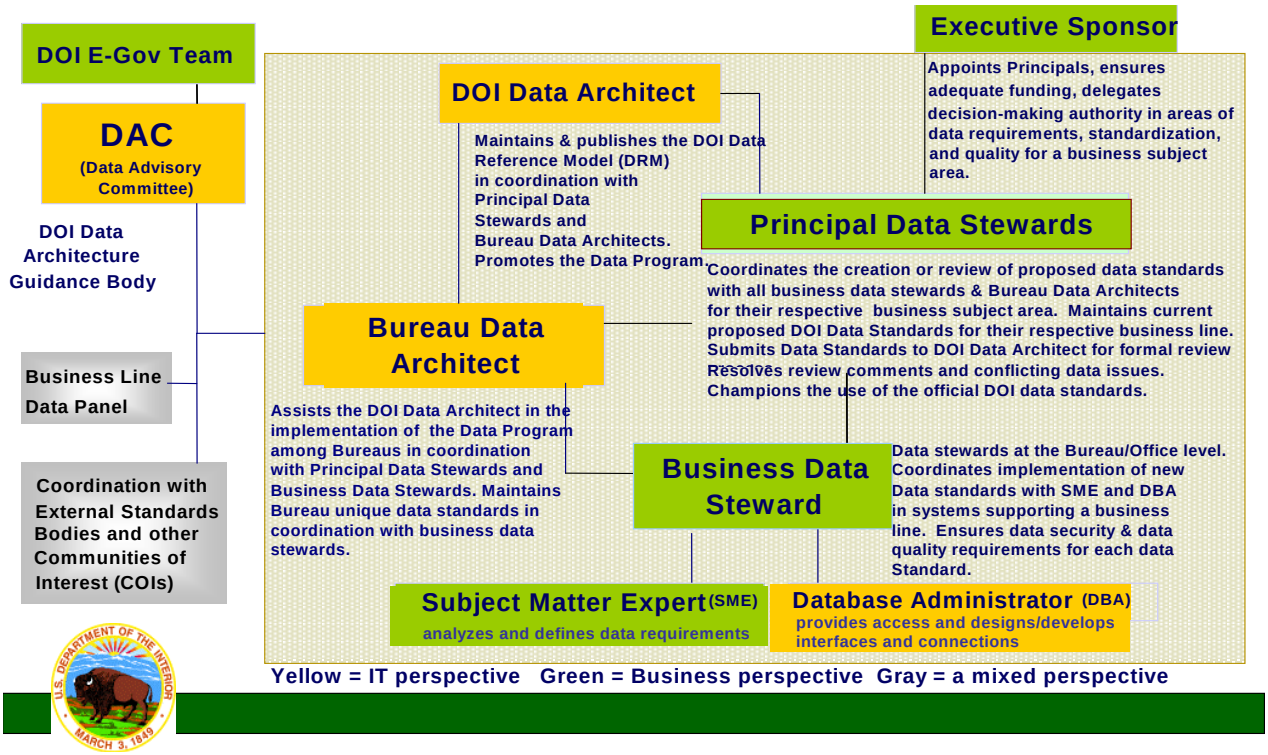

Figure 1-1: DOI's Data Governance Roles and Relationships

1.7 DOI's Data Quality Improvement Process Overview

This Guide describes the major activities that DOI's Bureaus are responsible for in the development and implementation of their Data Quality Improvement Processes. Bureaus will:

- Identify, prioritize, and assess areas of opportunity for increased data quality.
- Determine the most effective approach for improving processes to ensure that defective data are no longer produced.
- Correct existing defective data.
- Certify that the process and the data are in compliance with expected levels of quality or quality standards.

The Data Quality Improvement Process (shown in Figure 1-2) is based on accepted industry standards and incorporates project management and data quality management principles. The method is iterative and may be repeated until the data reach the appropriate quality levels. All four processes are mandatory only when the data being assessed are in an officially designated Authoritative Data Source (ADS). An ADS is a single, officially-designated source authorized to provide a type or many types of data and information that are trusted, timely, and secure on which lines of business rely. The intended outcome of ADS implementation is to provide data and information that are visible, accessible, understandable, and credible to information consumers, which include DOI's business users, DOI's information exchange partners, and IT applications and services.

Other non-ADS data sources that choose to implement these processes should follow the Assessment, Correction, and Improvement Processes, with the Certification Process being an optional process to implement. The following is a summary of each of the activities described in the Data Quality Improvement Process:

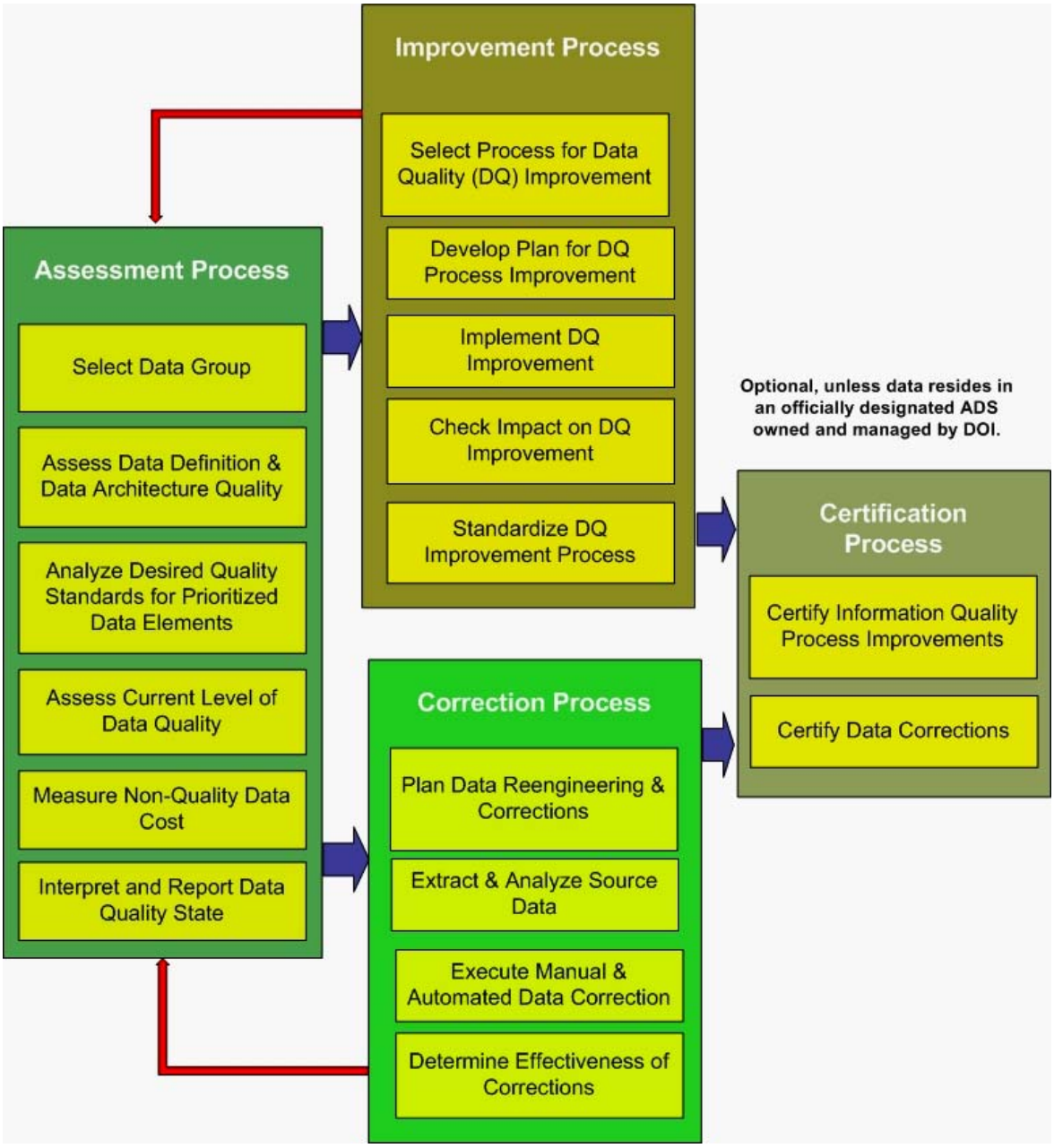

Figure 1-2: DOI's Data Quality Improvement Process

- **Implementing DOI's Data Quality Improvement Environment.** This chapter (Chapter 1) focuses on the systemic aspects that must be addressed within DOI to establish the proper environment for the successful deployment of a continuous data quality improvement process.

- **Data Quality Assessment Process**. Chapter 2 focuses on the assessment of the state of data quality. The Data Quality Project Team will execute this process in the assessment of mission-critical data. Data Quality Projects may apply these processes to internal data elements that are important to their functions and responsibilities. Assessment consists of selecting the information group candidates based on impact and priority, assessing the data definition and data architecture quality, determining the desired quality standards, assessing the current level of data quality, measuring the non-quality data costs, and interpreting and reporting the state of data quality. The outcome of the assessment, as a part of the final report, is a set of recommended follow-on actions for review by the OCIO and the Data Quality Project Team. The Principal Data Stewards and DOI Data Advisory Committee (DAC) are accountable for reviewing and accepting final assessment reports.
- **Data Quality Improvement Process**. Chapter 3 describes the activities that occur once the assessment has identified areas of improvement. At that point, the Data Quality Project Team will initiate activities to improve the quality of the data they acquire or produce. This is a proactive effort to minimize the incidence of defects in the data by attacking the causes of non-quality data. Improvement consists of selecting the process for data quality improvement, 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.
- **Data Quality Correction Process**. Chapter 4 describes the activities after the assessment process has identified areas in need of correction, during which the Data Quality Project Team will take steps to correct the quality of the data it acquires or produces. This is a reactive, one-time effort to eliminate existing defects in the data and should be taken as a complementary action to the improvement of the producing processes. Correction consists of planning the data correction, extracting and analyzing the data, executing manual and automated data corrections, and determining the effectiveness of the correction process.
- **Data Quality Certification Process**. Chapter 5 describes the activities of this optional task (except for ADS), which is one of independent verification or certification. This task takes place after the processes that produce or maintain selected data elements and information groups have been improved and after the existing data have been corrected. Based on the established priorities and schedules, Principal Data Stewards and the DAC will verify that the level of data quality achieved is aligned with the expectations of the business areas that consume the information.

Chapter 2. DATA QUALITY ASSESSMENT PROCESS

2.1 Overview

Assessment is the first step to achieve expected levels of data quality necessary for DOI to serve its constituents properly. Each Data Quality Project Team can use the following procedure to determine the data to assess:

- Conduct focus groups meetings 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 roadmap for improvement (i.e., Data Quality Plan).
- Select data (databases, files, other storage mechanisms) for assessment based upon the Data Quality Plan.
- Conduct time analysis for assessment.
- Perform a preliminary feasibility study of data repositories in scope to determine if complete baseline data (preferably three or more years old) are available for inspection.
- 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 that target compliance levels are maintained. The long-term goal is to achieve a state of continuous improvement, yielding the highest-quality data throughout the agency.

2.2 Select Information Group– Assessment Process Step 1

In this step, the Data Quality Project Team will select the data elements. With limited time and resources available, it is not feasible to correct every data element in every location and to analyze and improve every process that produced it. Therefore, the Data Quality Project Team will identify a set of criteria for selecting and prioritizing which data elements to assess. This is the process of determining the scope of a project.[6]

A. Determine Scope Based on Business Needs. The Data Quality Project Team will document the scope of effort to provide direction (e.g., set the parameters) for data quality improvement and data correction activities. To obtain the most value from these efforts, it is necessary to assess business needs, taking into consideration the entire data-value and cost chain that may be affected by data quality. To determine enterprise-wide business needs accurately, the Data Quality Project Team may conduct interviews with information consumers in each of the Bureau's business areas to ascertain their quality expectations by determining how they and their data stakeholders outside of DOI are using the data.

B. Identify Information Group to be Assessed. Once the business needs have been defined, the Data Quality Project Team will identify the data necessary to support those business needs. The Data Quality Project Team will collect the data from information consumer interviews in each of the Bureau's business areas to determine how they use the data in the performance of their jobs, and how the data support the needs of the business. Information on frequency of use of the data should also be collected. The objective is to determine the data for which assessment and

improvement processes could yield significant tangible benefits.[7] Table 2-1 illustrates how to document the data necessary to support the needs of the business.

Data Element Scope Worksheet *Example*				
Information Group	Data Element (Table or Record Name)	Within Scope? (Y/N)	Frequency of Use	Rationale for Inclusion or Exclusion (process and decision requiring it, and consequences if data are defective)
Inspection	Inspection Date	Y	Annual	For selecting inspections within the last fiscal year Printed on the report
	Inspection Report Completion Date	N	N/A	Not applicable to this report or indicator
Human Resource	Inspector First Name, Middle Initial, Last Name	N	N/A	Not applicable to this report or indicator
Property	Property ID	Y	Once	To distinguish a particular property in the computer system
	Property Name	Y	Once	Printed on the report
	Property Street Address, City, State, Zip	Y	Once	Printed on the report
	Property Contact Phone Number	N	Quarterly	Not applicable to this report or indicator
Organization	Regional DOI Office Street Address, City State, Zip	N	N/A	Not applicable to this report or indicator

Table 2-1: Illustration of a Data Element Scope Worksheet for a Fictitious Inspection Report

C. Identify Information Value and Cost Chain. The Data Quality Project Team will create a Value and Cost Chain (VCC). The VCC is critical because it will support future data analysis activities by aiding in the identification and understanding of the data flow across the enterprise.

There are several ways to diagram a VCC, and one of the most common ways is to create an Information Product (IP) Map. Figure 2-1 is a sample IP Map that shows the life cycle

movement of a single data element ('inspection_code') integral to the physical inspection of the business process. The IP Map:

- Identifies the files/databases that include the IP Map's attributes (i.e., data elements),
- Identifies the *database of or*igin 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 in which data are either duplicated or transformed,
- Facilitates the identification of stakeholders, and
- Can be leveraged for analysis of other IP Maps within a file/database.

Inspection_code is initially created by an Inspector on her or his palm device, shown in the upper left-hand corner of Figure 2-1. 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. 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 Recreation Information Database (RIDB), while the *database of record* is a cylinder outlined with a heavy, bold line (Recreation Reporting Database (RRDB). 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.)

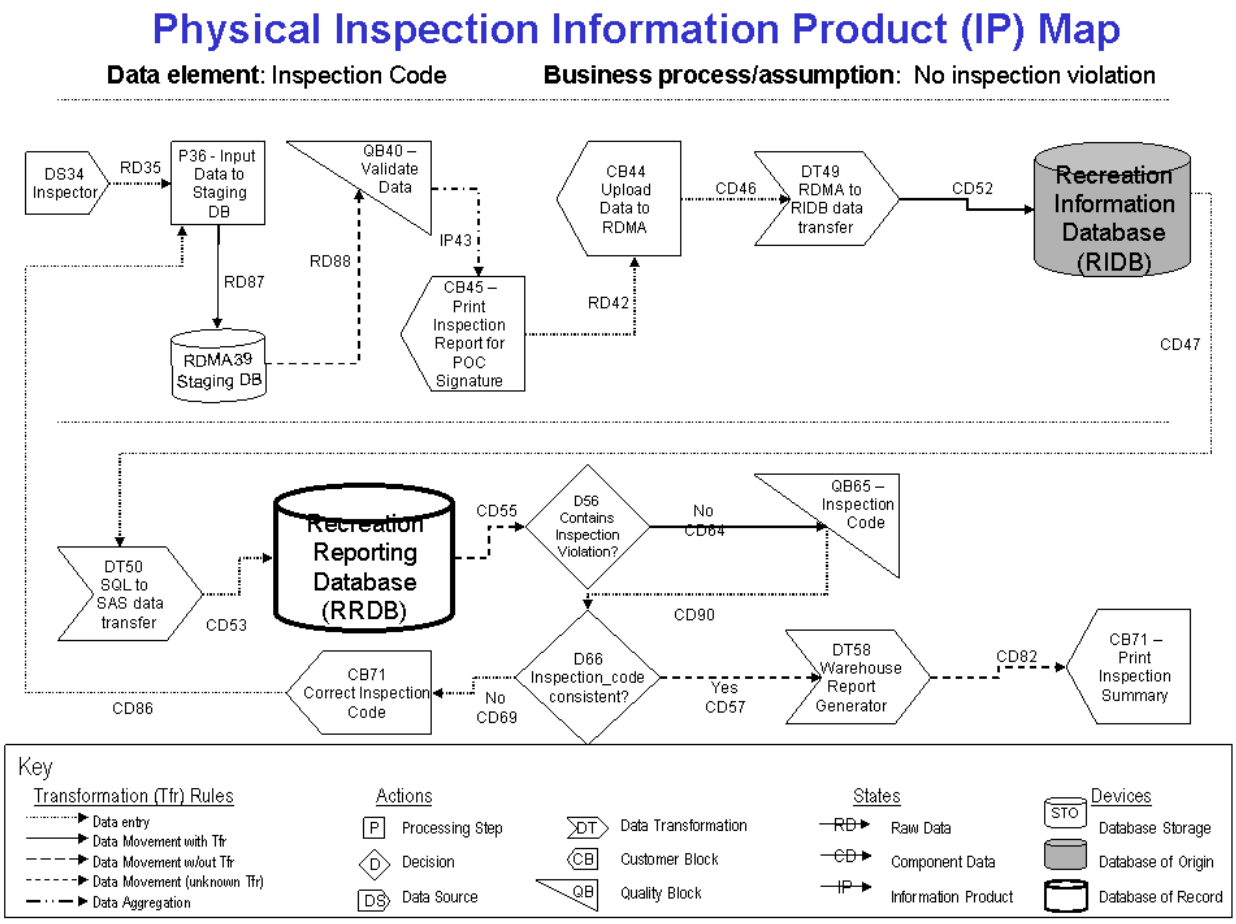

Figure 2-1: Sample Information Product (IP) Map

D. Identify Data Stakeholder. The Data Quality Project Team will identify the categories of data stakeholders. These stakeholders include:

- The data producers (including DOI's Bureaus, support offices, other federal agencies, and state and local governments) that create or maintain the data.
- The data consumers who use the data, including Data Quality Projects and support offices within DOI.
- The key contacts who maintain these data for future use.

E. Identify Data Content Quality Objectives and Measures. The Data Quality Project Team will establish the data quality dimensions to be measured in the information group to be assessed and will determine the quality objectives for each information group. These data can be gathered in the information consumer interviews and questionnaires. These measures may evolve while establishing quality standards. Table 2-2 describes the data quality dimensions.

Dimensions	Dimension Type	Quality Dimensions Description	Example of Non-Quality Data
Validity	Content	The degree to which the data conform to their definitions, 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	Content	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	Content	The degree to which the required data are known. This includes having the required data elements (the facts about the object or event), having the required records, and having the required values	An indicator for spouse is set to "yes", but spousal data are not present
Relationship Validity	Content	The degree to which related data conform to the associative business rules	A property address shows a Michigan zip code, but a Florida city and state
Consistency	Content	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 a different name or address or has different dependents
Concurrency	Content	The timing of updates to ensure that duplicate data stored in redundant files are equivalent. This is a measure of the data float (the time elapsed from the initial acquisition of the data in one file or table to the time they are 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 five days between the record-of-origin file and the record-of-reference database
Timeliness	Content	The degree to which data are 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

Dimensions	Dimension Type	Quality Dimensions Description	Example of Non-Quality Data
			data
Accurate (to reality)	Content	The degree to which data accurately reflect the real-world object or event being described	The home telephone number for a customer's record does not match the actual telephone number
Accurate (to surrogate source)	Content	The degree to which the data match the original source of data, such as a form, application, or other document	An applicant's income reported on the application form does not match what is in the database
Precision	Content	The degree to which data are known to the right level of detail (e.g., the right number of decimal digits to the right of the decimal point)	A measurement of water quality only recorded concentrations in parts per thousand whereas known contaminants can cause serious illness when found in concentrations of parts per million
Derivation Integrity	Content	The correctness with which derived data are calculated from their base data	The summary of accounts for a given district does not contain valid entries for the district

Table 2-2: Dimensions of Data Content Quality

F. Determine Files and Processes to Assess. Depending on the assessment objectives, the Data Quality Project Team will measure data at different points in the VCC. (See Table 2-3.) Data, such as record of origin and record of reference, are detailed in the files and databases within the VCC. In addition, the VCC identifies synchronized secondary stores, unsynchronized secondary stores, and yet-to-be categorized stores or applications.

The Data Quality Project Team will identify the system(s) that capture, maintain, or use the information group and assess the same data in the applications and files or databases. There is a tendency to assess the quality only in the circle of influence (e.g., the owned application or database); however, critical impacts on the Department can occur when the data are not of the expected quality and are shared across applications and Data Quality Projects.

As a general rule, 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 Data Quality Project Team's assessment must include the relevant databases, applications, files, and interfaces. In all cases, the approach to be taken must be defined and documented.

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.
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, extract records from both and 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 the 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 2-3: Data Quality Assessment Point by Assessment Objective

G. Prioritize Data Elements Supporting Business Need. When assessing and prioritizing the data elements, the Data Quality Project Team will consider the business needs across the *entire* data VCC, rather than solely in the system(s) identified as a Record of Origin. (Refer to Section 2.4 (B) of this Guide.)

The first step involved in this process requires the Data Quality Project Team to identify a list of data elements to be assessed. Once the data elements necessary to support the business needs have been identified, the Data Quality Project Team will prioritize the data elements. A simple high-medium-low scale may be used. Information consumers who best understand how the data meet their requirements make this determination. As stated in Section 515 Guidelines, "The more important the data, the higher the quality standards to which it should be held." Factors making a data element high priority might be:

- Importance to key decision making.
- Internal or external visibility.
- Impact on financial reporting.
- Operational impact of erroneous data (e.g., wasted time or resources).

Table 2-4 illustrates how to document the data element prioritization.

Data Element Prioritization Worksheet			
Data Class	**Data Element (Table or Record Name)**	**Data Element Priority (High, Medium, Low)**	**Rationale for Priority (process and decision requiring it, and consequences if data are defective)**
Inspection	Inspection Date	High	This is critical to determine if the inspection was performed within a year.
	Inspection Rating	High	This is critical to determine if the property passed inspection.
	Inspection Comments	Low	Not critical for this report.
Property	Property ID	High	This is the identifier of the property data in the computer system.
	Property Name	Medium	This is an important characteristic of the property but is not indispensable for the report.
	Property Street Address, City, State, Zip Code	Medium	This is an important characteristic of the property but is not indispensable for the report.
	Property Owner's First Name, Last Name	High	This is a determinant characteristic of ownership. It is required to assess if proper practices are in place.
	Property Owner Middle Initial	Low	Not critical for this report.

Table 2-4: Illustration of a Data Element Prioritization Worksheet for a Fictitious Inspection Report

2.3 Assess Data Definition and Data Architecture Quality – Assessment Process Step 2

The Data Quality Project Team will determine the quality measures for data definition and data architecture. The team will also evaluate the Logical Data Model for data structure (e.g., field size, type, and permitted value), relationships among the data, and the definition of the data under assessment. A logical data model is an abstract representation of the categories of data and their relationships. In this step, the Data Quality Project Team will develop or refine definitions and structures that are missing or have defective definitions or structures. In the case of defective definitions or structures, the Data Quality Project Team will recommend improvement in the data development and maintenance processes that created the defective definitions and architectures.

A. **Identify Data Definition Quality Measures.** The data definition must be consistent across the enterprise. Data needed to supplement or support analysis and make decisions should be

imported from ADS's and DOI's other Bureaus. The VCC diagram developed in Section 2.1 of this document identifies the files and databases in which each data element is stored. A comparison of the data definitions and storage format definitions across these files and databases is made to determine the level of consistency across the enterprise. Each definition is also assessed against the established Data Definition Quality Measures for compliance.

The Data Quality Project Team will identify and, if necessary, define the essential and critical quality dimensions of data definition and data architecture as defined in DOI's Data Standardization Procedures. These quality dimensions must be in place for ensuring effective communication among data producers, data consumers, or information consumers, as well as data resource management and application development personnel.[8]

In cases where no formal data definitions have been compiled or maintained, the Data Quality Project Team can derive partial 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 (e.g., forensic analysis).
- Identification of possible primary and foreign keys.
- Analysis of the database loading program for rules by which data columns are generated.
- Possible data formats.
- Data usage (e.g., column A is populated 3% of the time within Table B).
- 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 substantial domain or validity rules, a focus group consisting of the Data Quality Project Team, the data producers, and appropriate stakeholders will develop definitions to be used for assessment.

B. Assess Data Definition Technical Quality. The Data Quality Project Team will assess the data definition for conformance to DOI's Data Standardization Procedures determining whether the data definition conforms to the established minimum standards. The Data Quality Project Team will obtain a comprehensive and concise definition for each data element in the information group. This definition must contain an agreed-to statement or rule about the data content of the data element and its representation, the business rules that govern its data integrity, and the expected quality level based on the entire value chain of the data element.

C. Assess Data Architecture and Database Design Quality. Data Architecture and Database Design Quality are defined as the quality of the mechanism a data system employs for ensuring that data are well managed within the environment and distributed in an accurate, readable format within the system's repository and to other units within the organization based on business need. Well-designed data architectures allow disparate data to be captured and funneled into data that the business can interpret consistently for reporting results and to conduct planning appropriately for the future. Inadequately designed data architectures, which are not synchronized with the functional requirements of the business area or are not scalable to adapt to changing requirements, can lead to a misalignment between the technical implementation of the data flow and the demands made to use the data for reporting. This misalignment can impact the system's data quality findings and certification status.

The data architecture of the system is assessed against current data architecture best practices, with an eye to its alignment with the critical business performance indicators. The Data Quality

Project Team will assess the data architecture (or logical data model), the database design (implementation model), and the physical implementation of the data structures against modeling, design, and implementation of best practices by reviewing the following:

- Whether the data model has the required entity types and attributes to support the business processes.
- Whether the data model truly reflects the required business entity types, attributes, and relationships.
- Whether all instances of data redundancy in the system's data files, or other storage mechanisms, are controlled.

D. Assess Customer Satisfaction with Data Definition and Data Architecture. The Data Quality Project Team will measure customer satisfaction with the definition of the information products based on the information consumers' assessments. The deficiencies discovered in this step are critical input to the process discussed in Section 3.2. In this case, the processes that can be improved are the data definition and application development processes.

E. Develop or Improve Data Definitions and Data Architecture. In cases in which the data to be assessed in the subsequent steps lack or have defective definition and/or data architecture, the Data Quality Project Team will develop correct definitions and/or data architecture to ensure that subsequent tasks can be executed. The Data Quality Project Team will interact with representatives of the business and IT areas across the value chain to arrive at appropriate definitions and architecture. (See Section 2.2 E for a discussion of the identification of the stakeholders.)

To achieve a new or revised definition, the Data Quality Project Team will develop the necessary common terms and business concepts and then use them to define the entities, data elements, and relationships. The terms, entities, data attributes, and relationships must be coordinated and validated by the stakeholders across their value and cost chains.

F. Improve Data Development Process. If there is a pattern of missing or unsatisfactory data definitions that would be required in order to implement effective edit and validation routines or if data are defined with multiple meanings (e.g., overloaded data), then the data development and/or data maintenance processes are probably defective. If so, the Data Quality Project Team will recommend a process improvement initiative to improve the defective process (defined in Chapter 3). This improvement must be done prior to the next project that requires new data to be defined and implemented.

2.4 Analyze Desired Quality Standards for Prioritized Data Elements - Assessment Process Step 3

A. Define VCC Data for Data Elements. The Data Quality Project Team will perform in-depth analyses of the data elements that are within scope. These analyses should identify the data VCC of each data element or group of data elements, describe the element's quality dimensions, and determine the element's quality standard for each quality characteristic.

B. Define Record of Origin for Data Element(s). In order to assess data adequately, the Data Quality Project Team will identify the record(s) of origin for the data. Currently, there are cases in which data elements entered in an initial database are updated in a second, or even a third, system. In cases in which redundant data are identified, they must be corrected in every database in which they are stored. Table 2-5 is a template that should be completed for each system identified as a record of origin.

Data Element By Record of Origin System						
Data Element Business Name	Record of Origin System Name	Physical Data Element Name	Definition	Field Type	Length	Create/ Update
Inspection Date	DQ1	LAST-INSPECTED	The date the most recent property inspection took place.	Numeric	8	Create, Update
Inspection Rating	DQ1	INSPECTION-RATING	A classification indicating the relative condition of the property at the time of the inspection.	Numeric	3	Create, Update
Property Owner's First Name, Last Name	DQ1	OWNER-NAME	The First and Last Name of the person registered as legal owner of the property.	Alpha-numeric	40	Create, Update
	DQ2	OWNER-FORMAL-NAME	The fully formatted name of the owner. In the case of individuals, it is the combination of the First, Middle and Last Name. In the case of organizations, it is the legal name.	Alpha-numeric	50	Update

Table 2-5: Illustration of a Data Element by Record of Origin Worksheet

Once the record of origin has been identified, the Data Quality Project Team will determine where other read-only versions of the data are located in the organization, if possible. Strategies can then be formulated regarding the validation and correction of data at those data sites.

C. Identify Accuracy Verification Sources. In order to verify the accuracy of a data element value, the Data Quality Project Team will identify the most authoritative source from both surrogate and real-world sources from which to assess and confirm the accuracy or correctness of the data value. The most accurate of these is the real-world source, since the surrogate sources have a greater potential to contain errors.

D. Determine Applicable Quality Standard for Each Data Element. The Data Quality Project Team will determine and document the criteria for data quality according to the quality criteria discussed in Section 2.2 F. In establishing the desired level of data quality, the criteria should be assessed as to relevance and level of importance. Data that meet the criteria are considered quality; those data elements that do not meet the criteria are considered "defective" and must be corrected or discarded. Table 2-6 illustrates how to state data element quality criteria. It is important to understand that the accuracy criteria descriptions must explicitly name the data validity source that is the basis of the data in the record of origin. If no data validity source is available, then the method of determining accuracy must be described.

E. Determine Quality Compliance Levels. After defining the quality criteria for each data element, the Data Quality Project Team will determine what percentage of the data must comply with the specifications. This percentage will be the measure that is referenced when an

organization performs an internal quality assessment. Additionally, the Data Quality Project Team will use these percentages as the compliance levels when determining data element quality compliance.

The compliance percentage should be stated for each criteria specification. For example, a 100% Validity compliance target means that no data can deviate from the validity criteria. A 98% Accuracy level means that at least 98% of the data must meet the Accuracy criteria. If a data element meets *all* stated quality compliance targets, then the data element passes and is categorized as "quality compliant." Table 2-6 documents the data element compliance targets, as well as data exceptions.

Data Element Quality Criteria Worksheet for Record of Origin System: System X												
Data Entity / Data Element		Validity	Non-Redundancy	Completeness	Relationship Validity	Consistency	Concurrency	Timeliness	Accurate (to reality)	Precision	Accurate (to surrogate source)	Derivation Integrity
Inspection Date	**Quality Criteria**	Can be blank (not inspected) or a valid date since 1922.	-	-	With Inspection Rating: Either blank or both not blank.	-	-	-	-	-	Must match the inspection date on the inspector's log.	-
	Compliance level	100%	-	-	99%	-	-	-	-	-	99.5%	-
	Exceptions	-	-	-	-	-	-	-	-	-	-	-
	Findings	97% compliant; 3% are before 1922	-	-	4% missing when rating present.	-	-	-	-	-	5% did not match inspector's log.	-
Inspection Rating	**Quality Criteria**	Can be blank or numeric.	-	-	With Inspection Date: Either blank or both not blank.	-	-	-	-	-	Must match the inspection rating on the inspector's log.	-
	Compliance level	95%	-	-	100%	-	-	-	-	-	100% for 1996 and later.	-

Data Element Quality Criteria Worksheet for Record of Origin System: System X												
Data Entity / Data Element		Validity	Non-Redundancy	Completeness	Relationship Validity	Consistency	Concurrency	Timeliness	Accurate (to reality)	Precision	Accurate (to surrogate source)	Derivation Integrity
	Exceptions	-	-	-	-	-	-	-	-	-	Include only 1997 to current date.	-
	Findings	100% compliant.	-	-	4% present when date missing.	-	-	-	-	-	1% did not match inspector's log.	-
Property Owner's Name (First & Last)	**Quality Criteria**	Not blank. No special characters except hyphen, comma or period.	-	-	-	-	Must reflect changes received by DOI within 10 working days.	-	-	-	Must match owner's name in local authority's document of the assistance contract.	-
	Compliance level	99%	-	-	100%	-	-	-	-	-	98%	-
	Exceptions	-	-	-	-	-	-	-	-	-	-	-
	Findings	87% compliant; 13% are blank.	-	-	-	-	Current process takes up to 45 days to verify official change before updating the system.	-	-	-	Due to resource limitations, verified only lowest inspection ratings; found 22% names misspelled.	-

Table 2-6: Illustration of Quality Target Compliance for Record of Origin System

2.5 Assess Current Level of Data Quality - Assessment Process Step 4

A vital step in data quality improvement is to assess the *current level* of data quality. When selecting the data records for assessment, the Data Quality Project Team will acquire a representative, or statistically valid, sample to ensure that the assessment of the sample accurately reflects the state of the total data population, while minimizing the cost of the assessment. To be a statistically valid sample, "every record within the target population has an equal likelihood of being selected with equal probability."[9] When properly conducted, a random sample of records provides an accurate picture of the overall data quality of the database.[10] In certain circumstances, a selected sample of data may be usefully substituted for random samples. For example, if only active cases are of interest, the sample may include the active cases.

A. Measure Data Quality. The Data Quality Project Team will analyze the data in the samples against its target criteria based on the data definition and data architecture (defined in Section 2.3 F) and the defined quality standards (defined in Sections 2.4 D and E). The assessment should be performed against the established specifications, compliance targets, and data exceptions. Different data elements may require different assessment techniques for the various criteria. For each information group, either automated or physical data assessments or both are performed.

For accuracy assessment or certification, the authoritative source for the data element must be specified in the VCC. This may be a hard-copy document, a physical inspection of the real object or event the data represent (or a review of a recording of an event), data from an external source that is considered to be accurate, or an official document (e.g., a certified land survey) that is considered to be accurate.

B. Validate and Refine Data Definitions. The data definitions and architectures may be adjusted based on facts discovered during the measurement process. In such cases, the Data Quality Project Team will apply the data architecture and data definition process described in Section 2.3 F to arrive at the appropriate revised definitions.

C. Establish Statistical Control. For processes that acquire, produce, or maintain mission-critical data, it is imperative that they be in a state of statistical control. That is, with respect to the mission-critical data they produce, their results are predictable and the quality of the data is in line with the initially agreed-to levels of data quality.

2.6 Calculate Non-Quality Data Costs - Assessment Process Step 5

The objective of this step is to identify the cost of non-quality data for the information groups or data elements under assessment. Non-quality data costs are assessed in three areas: process failure costs, data scrap and rework costs, and lost or missed opportunity costs. *Process failure* costs are a result of a process, such as distribution of funds, which cannot be accomplished due to missing, inaccurate, incomplete, invalid, or otherwise non-quality data. *Data scrap and rework costs* are incurred when an information consumer has to spend non-productive time handling or reconciling redundant data, hunting for missing data, verifying data, or working around defective processes. *Lost or missed opportunity cost occurs* when DOI may be missing out on opportunities to greatly improve the services it provides due to non-quality data or may be directing funds toward services of lesser need.

A. Understand Business Performance Measures. Data have value to the extent that they enable the enterprise to accomplish its mission or business objectives. In order to determine if a process or data set adds value to the organization, it is important to understand the business vision and mission, the business plans and strategies, and the strategic business objectives.

B. Calculate Data Costs. The Data Quality Project Team will identify the percent of the system's data development and maintenance that adds value and the percent that adds cost, performed solely to correct the system's data quality.

2.7 Interpret and Report Data Quality State - Assessment Process Step 6

Once data have been assessed, the results will be analyzed, interpreted, and presented clearly in a format that can be easily understood by information consumers, data producers, and process owners. The results will include assessments of the components (e.g., definition and content). Also, the results will describe findings and recommendations in the quality standards (expected levels of quality), actual quality levels, and data costs, especially non-quality data costs. The Data Quality Assessment Report will include a cover sheet, a summary, a detailed section, and an assessment procedure report for each information group. (See Figure 2-2.)[11]

ASSESSMENT PROCEDURE REPORT

Information Group Name: _____
Time Period Covered From: _____ **To:** _____ **DQ Analyst(name):** _____
Sample Date:_____**File(s) Sampled:** _____**Processes Sampled:** _____
Sampling Procedure: □ Full Sample
□ Purposive Selection Sample (Purpose):_____
Sample Size: _____ **Sample Percent**: _____%

Assessment type: □ Electronic □ Third-party corroboration
□ Physical to surrogate source □ Survey
□ Physical to real object / event
□ Other: _____

Quality Dimensions Assessed:

□ Completeness of values
□ Reasonability tests / distribution analysis
□ Validity: conformance to business rules
□ Accuracy: correctness of values to: Source: _____
Surrogate: _____

□ Non-duplication of records
□ Timeliness of data availability
□ Equivalence and consistency of redundant data
□ Consistency
□ Concurrency
□ Precision
□ Derivation Integrity

Source: Improving Data Warehouse and Business Data Quality, Figure 6-9, 190, adapted for DOI.

Figure 2-2: Illustration of an Assessment Procedure Report Template

The Data Quality Project Team will perform a quality assessment and generate a Data Quality Assessment Report that describes the current level of data quality and makes recommendations for future project(s). These recommendations will address approaches for correcting the data errors identified and for changes to systems, procedures, training, and technology that will help to ensure the appropriate level of quality for the data. The final Data Quality Assessment Report will describe the following:

- The quality assessment approach, including the identification of the source system(s), the assessment criteria, and the specific DOI participants who are needed to conduct the assessment outside the membership of the Data Quality Project Team. These additional participants could be data producers, data consumers, system database administrators, and data architects.
- The current level of data quality, which includes general conclusions about the data quality of assessed elements, data quality defects found, and the assessment results (number and types of errors found, level of confidence in the results, and any other issues).
- The recommendations to close the gap between current data quality levels and target data quality standards. For data corrections, the recommendations should indicate the appropriate approaches for correcting the errors identified in the report. For data quality process improvements, the recommendation should identify the type of changes that may be made to systems, procedures, or technology to ensure the appropriate level of data quality. The recommendation should prioritize the list of tasks or areas of highest concern and indicate the anticipated start date.

[This Page Left Intentionally Blank]

Chapter 3. DATA QUALITY IMPROVEMENT PROCESS

3.1 Overview

Data quality improvement is a proactive step to prevent non-quality data from being entered into information systems. The data correction process corrects defective data, and this correction is part of the cost of non-quality data. The data quality improvement process attacks the *causes* of defective data. Eliminating the causes of defective data and the production of defective data will reduce the need to conduct further costly data correction activities.

Maintaining data quality is a continuing effort. Critical to the effectiveness of the procedure is a data quality awareness campaign that motivates data producers and information consumers to take daily ownership of data quality. As consumers and producers of quality data, information consumers and data providers are the best resources for identifying both quality issues and their solutions.

Data quality procedures must include periodic assessments to review data quality. This ongoing process ensures that the highest quality data are being used throughout the enterprise. When deficiencies in data are discovered, immediate steps must be taken to understand the problems that led to the deficiencies, to correct the data, and to fix the problem.

The improvement process consists of five major steps.[12] Those processes are:

- Select Process for Data Quality Improvement.
- Develop Plan for Data Quality Process Improvement.
- Implement Data Quality Improvement.
- Check Impact of Data Quality Improvement.
- Standardize Data Quality Improvement.

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

The implementation of data quality improvements will include one or more of the following actions:

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

3.2 Select Candidate Business Processes for Data Quality Improvement – Improvement Process Step 1

A. The first step in planning improvements is to identify which processes are the best candidates for process improvement. Candidate processes can be identified in the final Data Quality Assessment report outlined in Section 2.7. The candidate processes are then prioritized by the best return in data quality for the estimated investments in time, cost, and effort. The return on investment is estimated by reviewing:[13]

- Data Definition Technical Quality Assessment developed in Section 2.3.C.
- Data Architecture and Database Design Quality Assessment developed in Section 2.3.D.
- Level of Data Quality Assessment developed in Section 2.5.
- Cost of non-quality data calculated in Section 2.6.
- Any Customer Surveys that may have been conducted to determine data definitions.

B. Based on the nature of the problem(s) to be solved, a Quality Improvement Team (as defined by the Bureau) with representatives from all stages in the value chain is put into place. For a data quality improvement initiative to be effective:

- Quality Improvement Team representatives must perform the actual work.
- A non-blaming, non-judgmental environment for process improvement must be established. If there are defective data, it is because there are defective processes, not defective people.
- The process owner or supervisor of the process to be improved must empower the team to make and implement the improvements.
- The team must be trained in how to conduct a Root-Cause Analysis and how to learn what kinds of improvement and error-proofing techniques are available. The DAC must provide training materials.
- A process improvement facilitator must be available and trained in conducting the Shewhart[14] cycle of Plan, Do, Check, and Act (PDCA) process-improvement method.
- The origin of the data and their downstream uses must be understood.

3.3 Develop Plan for Data Quality Process Improvement – Improvement Process Step 2

The foundations for developing data quality procedures are the results obtained from the investigation of current processes controlling the data and the evaluation of possible root causes. Both manual and automated data control processes must be considered. The sources of the data must be considered, as well as who modifies the data or influences how the data are presented on a form, on a screen, or in a report.

A. Conduct Root-Cause Analysis. Once the process is understood, the Data Quality Project Team should analyze a data defect to identify the "root cause" of the defect using the Cause and Effect or Fishbone Diagram (Figure 3-1**),** the "why analysis" technique, or any other method for root-cause analysis. Six possible categories of failure causes are included in the Cause and Effect diagram, i.e., Human Resources, Material, Machine, Method, Measurement, and Environment. For each possible failure cause identified, it is necessary to answer "why" the error occurred until the root cause is found. The possible scenarios for tracking down the root cause should be explored by considering the six categories in the analysis. A typical root-cause analysis might be developed as:

Scenario - Defect identified is "Customer Number not on the Order":

- Why is the Customer Number not on the Order? Because the customer did not have the number (Material Cause).
- Why didn't the customer have the number? Because the customer has not yet received the mailing that contains the Customer Number (Method Cause).
- Why it was not supplied? Because the customer is new and ordered before receiving the Customer Number (Method Cause).
- Why did it cross in the mail? Because the new customer mailing runs only once a month (Method Cause).

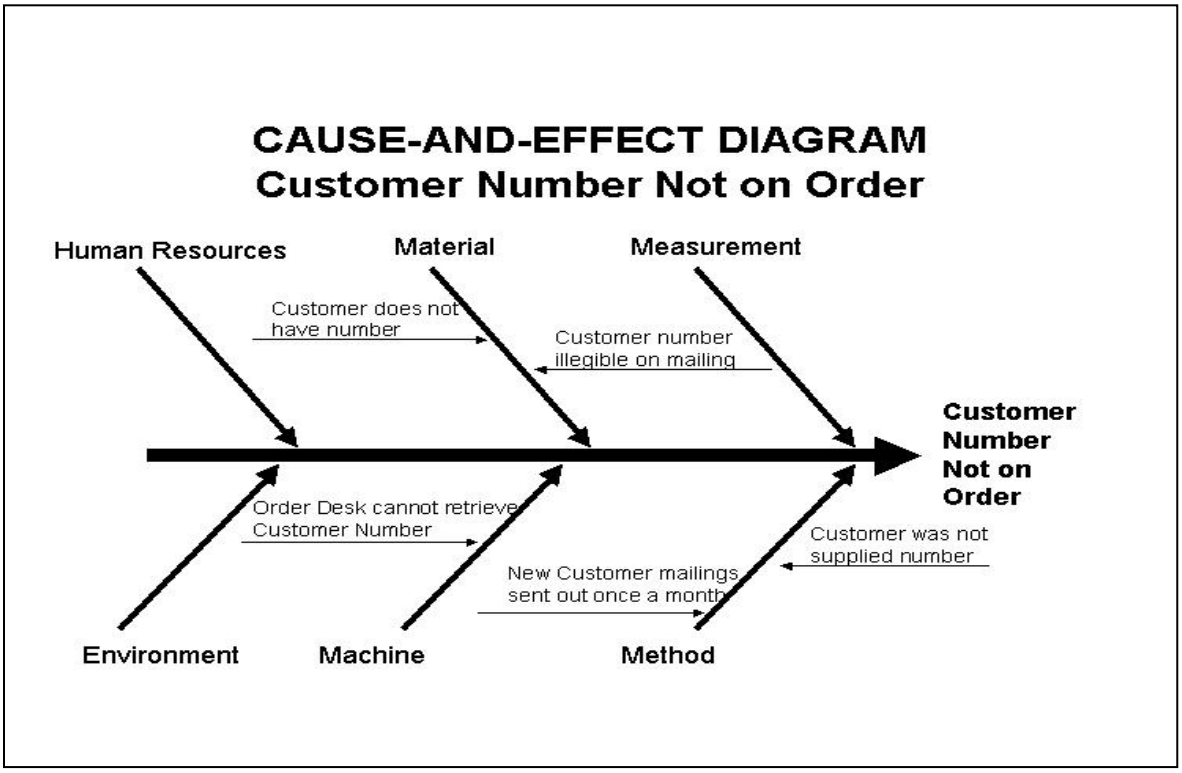

Figure 3-1: Illustration of a Cause-and-Effect Diagram

The diagram in Figure 3-2 presents typical areas of concern when applying the Cause-and-Effect diagram to the study of a Data Quality Issue.

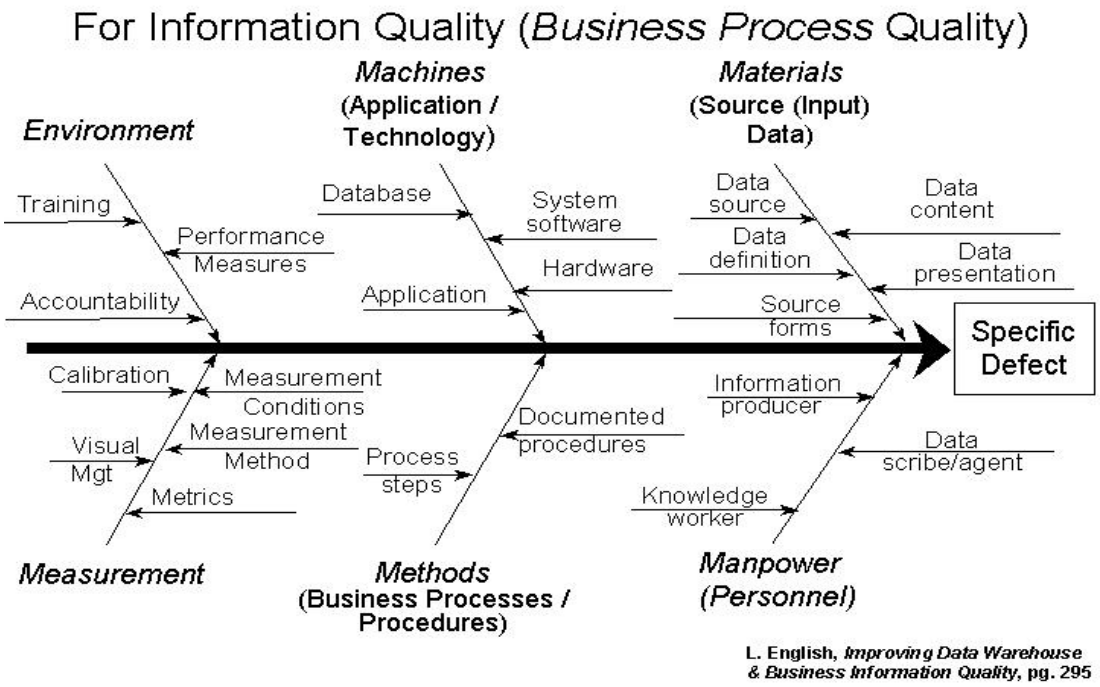

Figure 3-2: Cause-and-Effect Diagram Template for Data Quality

B. Define Process Improvement(s). The Data Quality Project Team will identify process improvement(s) only *after* root cause(s) have been identified and understood. Otherwise, it may be that only the symptoms or the precipitating causes of the problems are being attacked rather than the root causes. Data correction activities can also be leveraged into process improvements:

- Any automated correction techniques should become permanent changes in the software.
- Any manual correction procedures should either become permanent changes in the software or be transitioned to heavily-emphasized sections in the data quality awareness and training programs.

Just as the categories of failure may be Human Resources, Material, Machine, Method, Measurement, and Environment, the recommended improvements may involve any or all of these categories. Improvements should not be limited to "program fixes." Each category of failure requires a different type of improvement. Other categories may provide improvements that can be implemented easier, faster, or at lower cost. Examples of solutions from the categories include:

- Revise business processes to include procedures to ensure data quality. For example, include supervisory review of critical transactions before entry.
- Enforce data stewardship by holding managers and business process owners accountable, before data producers, for the quality, completeness, and timeliness of data, as well as for other required data quality standards.
- Allow data element domains to have a value of "unknown" in order to allow data producers to identify an unknown data value rather than entering a guess or leaving the entry blank. The data producer may not know all possible data values, and the "unknown" value may allow for future analysis and correct interpretation.

- Identify and designate official record-of-origin, record-of-reference, and authorized record-duplication databases.
- Adequately train data producers.
- Define data quality targets and measures and report data quality regularly.
- Implement effective edits that may prevent entry of defective data or flag entry of defective data for later correction.[15]

The Data Quality Project Team should link the specific set of data quality standards, procedures, and performance measures to each data-control process. Ideally, performance measures should encourage the data producers to create or maintain quality data. The Data Quality Project Team will measure the performance at the time of data capture. The Data Quality Project Team may want to implement a single process for data creation and maintenance along with a single application program for each data type, such as stakeholder, address, and property. These standards, commonly-defined processes, and applications should be implemented as early as possible in the life cycle of the data (or value chain). The databases and data elements must also be standardized to support the data requirements of the data customers.

3.4 Implement Data Quality Improvement – Improvement Process Step 3

A. The Data Quality Project Team will identify the data quality improvements and implement the recommended solution in a controlled fashion. The Team will document the new procedures, training, software modifications, data model changes, and database changes, as required. The Data Quality Project Team will also identify a controlled area in which to test the process improvements and implement the changes. If a "people" process is changed, an orientation and draft procedures will be provided. If software changes are made, the new version of the software will be deployed to a test area, and information will be provided to the consumers via training and/or draft documentation, if necessary.[16]

3.5 Evaluate Impact of Data Quality Improvement – Improvement Process Step 4

A. Once the process improvement has been deployed to the test environment, the Data Quality Project Team will evaluate the results of the improvement to verify that it accomplishes the desired quality improvement without creating new problems. The Data Quality Project Team will measure and quantify the benefits gained in the business performance measures; quantify economic gains and record lessons learned. If the desired results are achieved but new problems are introduced, the implementation of the quality improvement must be adjusted (defined in Section 3.4), or a different solution must be identified (defined in Section 3.3).

3.6 Standardize Data Quality Improvement – Improvement Process Step 5

A. Once the data quality improvement has been evaluated, the old, defective process can be replaced.[17] The Data Quality Project Team will:

- Roll out the improvements formally by formalizing the improved business procedures and documentation and implementing the software and database changes into production.
- Implement quality controls, as necessary.
- Communicate to the affected stakeholders.
- Document lessons learned, best practices, cost savings, opportunity gains realized, and process improvement history.

[This Page Left Intentionally Blank]

Chapter 4. DATA QUALITY CORRECTION PROCESS

4.1 Overview

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.[18]

Data Correction applies to a variety of efforts such as:

- Deployment of a data warehouse or operational data store, using Extract, Correction, Transformation, and Load (ECTL) techniques.
- Deployment or redeployment of a new operational application. (This situation is commonly known as a "conversion.")
- Correction of data in place of an existing operational application or decision support application. (This is called a "correction in-place.")

In the first two cases, the term "source" applies to the operational systems providing the data to the data warehouse or the operational data store, or to the legacy system being replaced by the new operational system. Also, in these cases, the term "target" applies to the data warehouse, the operational data store, or the new operational application. However, in the third case, the term "source" and the term "target" apply to the system being corrected. (In this case, the source and the target are the same.)

In the first two cases, the data correction efforts are almost always included in the overall plan of the data warehouse deployment (the ECTL task) or the new application deployment (the conversion and correction task). In the case of ECTL and data correction in-place, the task will correct the data and improve the process concurrently to prevent the production or acquisition of defective data. In the case of ECTL, there may be a gap between correction and improvement due to resource constraints. Therefore, the defects identified by the ECTL components must be corrected, captured, and reported to the producing area. This applies whether the correction is one-time files (e.g., for historic files) or ongoing files with reference or transaction data provided by the operational systems.

For in-place corrections, it is important that there not be a time gap between data correction and implementing data quality improvements. Data correction and process improvement implementation should be closely coordinated to prevent additional correction of the same data in subsequent efforts.

4.2 Plan Data Correction – Correction Process Step 1

A. Conduct Planning Activities. The Data Quality Project Team will establish interim and completion milestones for each task to provide clear indicators of progress and problems. Careful planning shortens the time it takes to perform correction and will ensure that resources are available when needed. Several planning activities should occur in parallel:

- Determine the appropriate correction approach.
- Update the correction plan and schedule.
- Determine automated tool-support requirements and schedule.

B. Produce a Data Element Correction Plan. The Data Quality Project Team will create a Data Element Correction plan that includes the following information:

- Identification of correction steps for each data element/information group.
- Discussion of the feasibility of data element correction (i.e., are source documents available? Is it too costly to correct? Is a "correct" data element critical to the conduct of business within DOI or by any external partners?).
- Description of overall correction approach.
- Updated Work Breakdown Structure including tasks that identify the resources for data element correction and the automated tool-support requirements and schedule (e.g., required deliverable).
- A list of required deliverables.
- The updated, detailed correction schedule.

C. Identify and Prioritize Data to be Corrected. Using the data VCC (defined in Section 2.2 C) and the Data Quality Assessment Report (defined in Section 2.7) in conjunction with the Non-Quality Data Costs analysis (defined in Section 2.6), the Data Quality Project Team will rank the data by quality, cost to correct, and benefits if corrected. The state of quality, feasibility, and cost of correcting must be considered in designing the correction steps for specific data element(s).

D. Identify Methods for Data Correction. The Data Quality Assessment Report (defined in Section 2.7) provides a measurement of where and how each data element falls below the desired level of quality. Different quality defects may require different correction techniques:

- Identification and consolidation of duplicate data.
- Correction of erroneous data values.
- Supplying of missing data values.
- Calculation or recalculation of derived or summary data values.

The Data Quality Project Team will develop a set of corrective steps to reflect business rules affecting each data element. These steps are applied either manually or through automation to correct the data. In addition, the Data Quality Project Team will document a summary of the data defects and the related correction techniques/steps to be applied.

Once the appropriate correction steps for each data element or group of similar data elements have been documented, the Data Quality Project Team will describe the overall correction approach and finalize the schedule of resources and tasks. The schedule must be sufficiently detailed to include task milestones so correction progress can be readily monitored. Correction should be automated to the greatest extent possible to help eliminate errors. The lead time required for possible acquisition of tools/techniques and their associated training, development, testing, and production use should also be considered.

4.3 Extract and Analyze Source Data – Correction Process Step 2

Although the initial assessments detailed in Section 2.5 provide a measure of data quality, there may be "hidden" data stored in data elements that are not part of their formal definition. It is important that the data be examined to uncover anomalies and to determine whether additional data elements can be identified. The Data Quality Project Team will analyze and map the data against the data architecture (defined in Sections 2.3 and 2.5) to ensure that the data elements are identified and fully defined with the associated business rules.

A. Plan and Execute Data Extraction. The Data Quality Project Team will perform a random sampling of data that are extracted from the source database or a set of related databases. (Refer to Section 2.5 B.) Any method may be used to generate the random sampling, as long as a fully representative sample is produced.

B. Analyze Extracted Data. The Data Quality Project Team will parse the extracted data into the atomic-level attributes to ensure that the data are examined at the same level. Once parsed, the specific data values will be verified against the data definition to identify anomalies. The data will be reviewed by subject matter experts to confirm business rules and domain sets and to define revealed "hidden" data (i.e., data whose structures are not found in a data dictionary). Also, the data will be reviewed for patterns that may reveal not-yet-documented business rules, which will then be confirmed by the subject matter experts. It is not unusual to find that data that first appeared to be anomalous can help rediscover forgotten business rules.

C. Document Findings. The Data Quality Project Team will document the definition, domain value sets, and business rules for each data attribute in the database or set of related databases that are documented in the Data Definition Worksheet. (See Figure 4-1.) The relationship of the data attributes is mapped to the source files and fields using the Data Mapping Worksheet. These data will be used in the transformation process.

Data Definition Worksheet

System: SAMPLES

Data Element Storage Details:

Table: Voucher **Column:** Contract Number

Storage Format: Text **Length:** 10

Definition: The contract number is a unique identifier issued upon contract initiation for Section 8, Section 202 PRAC, Section 811 PRAC, and Section 202/162 PAC subsidy contracts.

Domain Values: N/A

Business Rules:

1 Value contains letters and numbers only.

2. If value begins with a letter, then value must be a two-letter combination that corresponds to a valid state code.

3. If the subsidy type is 1, 7, 8, or 9, then a value must be present.

4. If the subsidy type is 2, 3, 4, or 5, then a value must NOT be present.

Figure 4-1: Illustration of a Data Definition Worksheet

Data Mapping Worksheet		
Data Element	**RIDB Head of Household SSN**	**RRDB Head of Household ID**
Definition	Social security number of the head of household is the unique identifier of a family.	Head of household id is a unique identifier for households receiving housing assistance. It is either the head of household's social security number or a system generated ID beginning with "T."
Storage Format	Numeric length 9.	Text length 9.
Domain Value Sets	N/A	N/A
Business Rules	SSN of head of household must be numeric.	Value must contain a 9-digit number or the letter "T" followed by 8 digits.
	SSN of head of household must be 9 digits.	
		Value cannot start with the number "9."
		Value cannot start with the number "8."
		Value of the first three digits cannot be "000."
		Value of the middle two digits cannot be "00."
		Value of the last four digits cannot be "0000."
		Value of the first three digits cannot fall between 766 and 799.
		Value of the first three digits cannot fall between 729 and 763.
		Value of the first three digits cannot fall between 681 and 699.
		Value of the first three digits cannot fall between 676 and 679.
	SSN of head of household must	Value cannot be "000000000."

Data Mapping Worksheet		
Data Element	**RIDB Head of Household SSN**	**RRDB Head of Household ID**
	not contain a suspicious value, such as 000000000, 111111111, 222222222, 333333333, 444444444, 555555555, 666666666, 777777777, 888888888, 999999999, 123456789, and 987654321.	Value cannot be equal to 111111111, 222222222, 333333333, 444444444, 555555555, 666666666, 777777777, 888888888, 999999999,123456789, and 987654321.
		Value must be unique with certification effective date and change sequence number except for the case of 999999999.
		Value is not null or blank.

Table 4-1: Illustration of a Data Mapping Worksheet

4.4 Execute Manual and Automated Data Correction – Correction Process Step 3

In this step, the manual and automated corrections are developed, tested, and executed. The corrections may be applied in-place, to an intermediary database, or to another target, such as a data warehouse or data mart. The basic techniques remain the same. Documenting the successes and missteps as they occur will enable the use of these correction techniques in subsequent projects.

A. Standardize Data for Atomic-Level Format and Values. The Data Quality Project Team will examine the data across databases for consistency in their definitions, domain values, and storage formats, use of non-atomic data values, and instances of domain duplicate values (e.g., Sept and Sep). If the data definitions and architectures require refinement based on the actual data in the files and databases, the Data Quality Project Team will initiate a data definition effort based on the process described in 2.3 F. Once the rules for standardization have been reaffirmed, the Data Quality Project Team will map the source data against the standardization and the data merge and transformation rules.[19]

B. Correct and Complete Data. The Data Quality Project Team will correct and complete the data identified in Section 4.2 to the highest quality that is feasible. This process is particularly significant if the source data are subsequently transformed and enhanced to be incorporated into a data warehouse or data mart. Data anomalies may include:

- Missing data values.
- Invalid data values (out of range or outside of domain value sets).
- Data that violate business rules such as invalid data pairs (e.g., a Retirement Date for an Active employee) or superfluous data (e.g., an Employee has two Spouses).
- "Suspect data," such as duplicate data values when unique values are expected; overabundance of a value; or data that "look wrong" (e.g., an SSN of 111-11-1111 or a Start Date of Jan 01, 1900).

Occasionally, some data may be "uncorrectable." The Data Quality Project Team may choose to address the situation by:

- Reject the data and exclude it from the data source.
- Accept the data as is and document the anomaly.
- Set the data to the default value or an "unable to convert" value.
- Estimate the data.

Estimating the data may be an acceptable solution, but the risk of using incorrect data should be carefully weighed. An estimated data value is, by nature, not correct, and incorrect data are often more costly than missing data.

The Data Quality Project Team will document the method for correcting each data element type and the method used for handling uncorrectable data (Figure 4-2**).** In addition, the Data Quality Project Team will document the cost for correcting each data type to track the expense of data cost and rework. Costs include:

- Time to develop transformation routines.
- Cost of data correction software.
- Time spent investigating and correcting data values.
- Cost of computer time.
- Cost of materials required to validate data.

Other costs associated with the non-quality data must be identified and quantified.[20] Non-quality costs include:

- Costs of non-quality data (scrap and rework) including: non-recoverable costs due to non-quality data; redundant data handling and support costs; business scrap and rework costs; work-around costs and decreased productivity; costs of searching for missing data; costs of recovery from process failure; other data verification/cleanup/correction costs; system requirements design and programming errors; software "re-write" costs; liability/exposure costs; recovery from process failure; and recovery costs of unhappy customers.
- "Losses," measured in revenue, profit, or customer lifetime value, including lost opportunity costs and missed opportunity costs.
- Mission failure (Risk) with impact, such as the inability to accomplish mission or even to go out of business.

Data Correction Worksheet

System: _____

Information Group (data element list): _____

_____

Data Correction or Uncorrectable Data Handling Method Used: _____

Expenses: _____

Time Investigating Data Defects: Man Days/Months/Years @ $ _____avg. cost

GOTS/COTS Data Correction Software Cost: _____

Time Spent Correcting Data Values: _____ Man Days/Months/ Years @ $ _____avg. cost

Time to Develop Transformation Routines: _____ Man Days/Months/ Years @ $ _____avg. cost

Cost of Computer Time: _____

Cost of Materials to Validate Data: _____

Total Costs: _____

Figure 4-2: Illustration of a Data Correction Worksheet Template

C. Match and Consolidate Data. In the cases where there is a *potential* for duplicate records within a single data source or across multiple data sources, candidates for possible consolidation are identified based on match criteria that meet the expectations of the stakeholders. Improperly merged records can create significant process failures and are therefore less desirable than duplicate records. Match criteria for merging records must be validated to ensure that duplicates are eliminated without creating improper merges.

The Data Quality Project Team will develop match criteria for more than one data element, with relative weights assigned to each match. If the impact of two incorrectly merged records is high, the match criteria should be rigorous. Examples of match criteria and relative weights/points are:

- Exact match on Name, 50% or 20 points.
- Phonetic match on Name, 35% or 15 points.
- Exact match on Address, 25% or 10 points.
- Close match on Address, 15% or 5 points.
- "Keyword" matches, such as Bob and Robert or Education and Training, 25% or 10 points.

Match criteria results are additive. In the example above, an exact match on Name and Address would yield a relative weight of 75% or 30 points, while a phonetic match on Name and close match on Address would yield a relative weight of 50% or 20 points.

The Data Quality Project Team will examine the records with matches to determine if they are indeed duplicates. If the duplicates can be traced back to two different data sources, the records should be cross-referenced in a control file to avoid the creation of duplicate records in the future. Consolidations of particular data types in specific data sources may be disallowed in some circumstances (e.g., if the records involved have been designated as Master Records and cannot be removed).

D. Analyze Defect Types. The Data Quality Project Team will analyze errors in the previous steps for patterns, costs, and impacts on the business. The patterns help identify problems, often

pointing to the source process. The costs and impacts help prioritize the possible process problems to be resolved. The Data Quality Project Team will combine and document the results in the Data Element Correction Summary Report with the following outline:

- Description of manual and/or automated correction tools and techniques used during data element correction.
- List of data files, records, and elements corrected.
- Updated Data Element Quality Criteria Specification Worksheet.
- Correction directives sent to headquarters and/or field staff.

E. Transform and Enhance Data. Once the data have been corrected, the Data Quality Project Team will prepare the data for loading back to the source database or into the target database. In the cases in which data transformation is required, the transformation process addresses any data conversions necessary as identified in Section 4.4 A. The enhancement process augments internal data with data from an external data source. The standardization rules applied to the data define the data transformation rules, and the data transformation rules are used to develop the transformation routines. Examples of expected data transformations include the following:

- **Data Extraction:** Selected fields are mapped to the target without conversion. For example, the Order database may include Order Number, Customer ID, Ship-to Address and Billing Address, while the target data warehouse database may require Customer ID and Ship-to Address.
- **Domain Value Conversion**: Non-standard domain values are converted to standard. For example, if the corporate standard is to use three character codes for month values, a database that stores months using the numbers 1-12 must be converted to the three-character code.
- **Codify or Classify Textual Data**: Free text data are converted to discrete codes or domain values. A common example of this is a "reason" text field, in which an examination of the data would yield candidate codes or domain values. Once converted to discrete codes or values, the data can be used statistically.
- **Vertical Filter:** A field used for multiple purposes is split into discrete fields for each purpose.
- **Horizontal Filter**: A field is split into atomic-level components. A common example of this transformation is splitting full name into first name, last name, and middle initial.
- **Matching and Consolidation**: Records identified in Section 4.4 C, above, and verified as true duplicates are consolidated.
- **Data Evaluation and Selection:** As records are combined from multiple data sources to a data warehouse or other database, the most authoritative data are selected. If in doubt, an informal quality assessment similar to the one performed in Section 2.5 can help identify the most correct source.

Enhancements include the addition of geographic, demographic, behavioral, and census data from an external source to support an identified business need. For example, income data may be obtained from an external source and appended to clients' records to help determine their Section 8 benefits.

F. Calculate Derived and Summary Data. If data are summarized or derived, the Data Quality Project Team will calculate these data. This usually applies to a data warehouse or data mart

ECTL. Data are summarized or combined to optimize performance for frequent queries made to the database. This can be accomplished through the following steps:

- The queries requiring the summary or derived data are identified.
- The calculation rules and/or algorithms supporting the queries are defined and verified with the SME or business data steward.
- The software routines for the derivation or summarization are developed and certified.

4.5 Determine Adequacy of Correction – Correction Process Step 4

Before the project can be closed, the Data Quality Project Team must evaluate the success of the correction process. At a minimum, the following steps are executed:

A. Perform post-correction quality assessment. The Data Quality Project Team will determine the compliance level of each data element's post-correction quality. This compliance determination ensures that: (1) the data values fall within the domain value set or range; (2) any "missing" data values are now present; (3) the data values follow business rules; and (4) the data are loading according to specified data mapping, as developed in Section 4.3 C.

The Data Quality Project Team will verify effects of transformation and enhancement. This verification ensures that: (1) the data are transformed as expected and (2) the records are enhanced with the correct data as expected.

The Data Quality Project Team will verify that the records are loaded as expected by: (1) ensuring that the jobs ran to completion; (2) validating that the correct number of records was processed; (3) validating that none of the records were inadvertently processed twice; and (4) ensuring that the correct number of duplicate records was consolidated.

B. Assess impact of data correction techniques. The Data Quality Project Team will document the impact of the correction techniques as percent of errors or omissions:

- Corrected accurately using automated means.
- Corrected through human efforts or means.
- Corrected to an inaccurate value (valid, but not accurate).
- Not corrected because it was impossible or cost prohibitive to get the correct value.

C. Recommend data correction improvements. The Data Quality Project Team will analyze the data defects, recommend appropriate improvements, and update the Data Element Quality Criteria Worksheet (Table 2-6) with the corrected results.

D. Document post-correction data correction findings. The Data Quality Project Team will document:

- The correction techniques that worked and that did not work.
- Adjustments to the correction schedule.
- Data element post-correction compliance levels.
- Analysis of data quality weaknesses and recommendations for corresponding improvements.
- Assessment of the correction plan, schedule, required human resources, and roles
- The improvement of data quality.

Chapter 5. DATA QUALITY CERTIFICATION PROCESS

5.1 Overview

This chapter describes the methods and techniques that will be used by the Data Quality Project teams to perform the final task of independent verification or certification of mission-critical information. This is an optional, but recommended, task, except for ADS. This task of independent verification, or certification, is conducted after the processes that produce or maintain selected data elements and information groups are improved and the existing data have been corrected. This certification will be in two areas:

- First, to assess whether the data produced by create and maintain processes are in compliance with the definition and quality standards of the information. This assessment will help evaluate and improve the effectiveness of process improvement efforts.
- Second, to assess whether the data contained in files, databases, data warehouses, data marts, reports, and screens are in compliance. This assessment will help evaluate the adequacy of data correction efforts.

Based on its observations and findings, the Data Quality Project Team will recommend improvements to the procedures used to implement data quality improvements (defect prevention) and improvements in the data correction procedures. Based on the established priorities and schedules, the Principal Data Stewards and the DAC will verify that the level of data quality achieved is aligned with the expectations of the business areas that consume the information. In addition, if the certification process finds shortfalls in information quality, the responsible Data Quality Project Team must submit a new schedule and perform additional information improvement and/or corrections.

5.2 Certify Information Quality Process Improvements – Certification Process Step 1

This activity is similar to the "Evaluate Impact of Data Quality Improvement" activity described in Section 3.5. Before a meaningful certification of an information process improvement can be performed, the process must be certified as being in statistical control. That is, the process must be producing a consistent (i.e., predictable) and acceptable level of quality of information (i.e., the data consistently meet the information consumers' and end customers' needs). Once the process is in statistical control, it is possible to determine that the changes indeed produced the expected improvements. The Data Quality Project Team will verify the effectiveness of the Data Quality Improvement process by assessing the results of the data quality improvement. Critical points to be assessed include:

- Was the data quality improvement planned appropriately? Is there anything that can be done to improve the process? The plans (i.e., the "P" in the PDCA cycle) and any documentation of the actual execution of the plans will be used to determine if the process must be revised for improvement.
- Was the data quality improvement implemented in a controlled environment? Was the control environment representative of the target environment? This is the process of determining the effectiveness of the execution (i.e., the "D" in the PDCA cycle).
- Were the data quality improvement results checked for impacts across the information value chain? This is the process of determining the effectiveness of the "check" (i.e., the "C" in the PDCA cycle).

- Were the actions to standardize the data quality improvement across the target environment effective? Were the expected results achieved? The actual rollout or "Act" (i.e., the "A" in the PDCA cycle) logs will be used to determine if "unplanned" events or activities can be prevented or mitigated in future efforts.
- If the Data Quality Project Team identifies a need for improvement in any of these areas, it will determine the root cause. This may necessitate application of the "why" technique or the fish-bone technique, as described in Section 3.3 B.

5.3 Certify Data Corrections – Certification Process Step 2

This section outlines the steps necessary to assess the adequacy of the Data Correction efforts to revise or correct existing data.

A. Define the Scope of the Certification. The Data Quality Project Team will identify the information group to be certified and the assessment points (files, databases, screens, reports) within their value and cost chain, using the same criteria as stated in Section 2.2 G but only for information groups that the Data Quality Project has identified as ready for certification and for the assessed data quality objectives and measures. This will produce a Scope Statement and Work Plan. The work plan is based on the original assessment plan. The Work Plan specifies the information group, the entire value chain, the operational systems, system interfaces, and analytical systems that will be certified, as well as the tasks to be conducted, dependencies, sequence, time frames, milestones, expected outcomes (products), and estimated time to complete. The plan will specify any assumptions, critical success factors, or risks.

B. Identify the Data Element Definitions. If the Data Quality Project Team has applied the Data Quality Improvement Process approach described in this Guide, the comprehensive definition will already be specified for each data element. Refer to Section 2.3 for a detailed discussion of this task. However, if the data definition is not in place, the Data Quality Project Team will develop data element definitions using the approach described in Section 2.3 F.

C. Define Certification Approach. Based on the prior assessment for each information group, the Data Quality Project Team will determine one or more techniques for assessing its actual level of quality. Refer to Section 2.5 C for details on this selection.

D. Define Sample Size and Resources. Based on the prior assessment, for each information group and for each assessment point, the Data Quality Project Team will determine the sample size using the same approach as the prior assessment. In addition, the Data Quality Project Team will identify the participants in the assessment process and the estimated number of hours and calendar days required and any special requirements, such as access to documents, acquisition of tool(s) not already in DOI's inventory, and travel requirement.

E. Develop Certification Criteria. The terms of the certification criteria will be the same as those agreed to as part of the original assessment data quality criteria levels, unless otherwise agreed to by the OCIO and the Data Quality Project team based on lessons learned during the data correction process or special conditions identified by either party.

F. Conduct Certification. The Data Quality Project Team will perform the tasks in the certification Work Plan to determine the level of compliance of the data elements within the scope of the certification.

G. Interpret and Report Information Quality Certification. Once the data have been certified, the Data Quality Project Team will report the results as stated in Section 2.7, replacing the term "assessment" with the term "certification."

APPENDIX A. DATA QUALITY IMPROVEMENT PROCESS PLANNING

DOI's Data Quality Improvement Process method defines four major processes. However, data quality assessment or certification projects frequently will add or remove steps or tasks within processes to meet the particular needs of the Bureau.

The decision concerning which processes will be conducted and the specific tasks to be performed in each step must be documented in a project plan. Although the entire project should be included in the plan, it will be necessary to update the plan at key points throughout the project. The level of detail in the plan will vary at different stages.

Samples of work breakdown structures are provided as a starting place in the subsequent sections of this appendix, to be tailored as needed for specific projects. The Data Quality Improvement Process tasks may be iterative based upon the requirements of the individual Data Quality Projects and the state of data quality. These samples were developed to help identify the high-level tasks required to plan and execute a Data Quality Improvement Process project. Additional tasks will be needed, such as training in the method at the beginning of the project, details of the assessment and correction processes depending upon the specific data elements and systems in the scope, and details of the improvements process, once specific improvements are identified.

A.1 OUTLINE FOR DATA QUALITY IMPROVEMENT PROCESS

A project plan typically includes the components listed below. The Data Quality Improvement Process Project Plan for an Improvement or a Correction project is a required document. However, only the project Schedule is a required deliverable.[21]

- Executive Summary: Describes the purpose, scope of activities, and intended audience of the plan.
- Project Objectives: Describes the business goals and priorities for management of the project.
- Project Assumptions, Constraints, and Risks: States the assumptions upon which the project is based, including the external events the project is dependent upon and the constraints under which the project is to be conducted. Identifies and assesses the risk factors associated with the project and proposes mitigation of the risks.
- Work Breakdown Structure: Identifies high-level tasks required for planning and executing the project.
- Project Responsibilities: Identifies each major project function and activity and names the responsible individuals.
- Task Descriptions: Describes each function, activity, or task and states both internal and external dependencies.
- Project Deliverables: Lists the items to be delivered and the delivery dates.
- Resource Requirements and Plan: Specifies the number and types of personnel required to conduct the project. Includes required skill levels, start times, and plans for training personnel in the Data Quality Improvement Process method. Includes requirements for computer resources, support software, computer and network hardware, office facilities, and maintenance requirements.

- Schedule: Provides the schedule for the various project functions, activities, and tasks including dependencies and milestone dates. A Gantt chart noting major deliverables and milestones is very useful to depict a summary view of the entire project schedule.

A.2 SAMPLE DATA QUALITY IMPROVEMENT PROCESS ASSESSMENT WORK BREAKDOWN STRUCTURE

1.0 Plan Data Quality Improvement Process Assessment Project

2.0 Select Information Group

- 2.1 Determine Scope Based on Business Needs
- 2.2 Identify Information Group to be Assessed
- 2.3 Identify Data Value and Cost Chain
- 2.4 Identify Data Stakeholders
- 2.5 Identify Data Quality Objectives and Measures
- 2.6 Determine Files and Processes to Assess
- 2.7 Prioritize Data Elements Supporting Business Need

3.0 Assess Data Definition and Data Architecture Quality

- 3.1 Identify Data Definition Quality Measures
- 3.2 Assess Data Definition Technical Quality
- 3.3 Assess Data Architecture and Database Design Quality
- 3.4 Assess Customer Satisfaction with Data Definition and Data Architecture
- 3.5 Develop or Improve Data Definitions
- 3.6 Improve Data Development Process

4.0 Analyze Desired Quality Standards for Prioritized Data Elements

- 4.1 Define Data Value and Cost Chain for Data Element(s)
- 4.2 Define Record of Origin for Data Element(s)
- 4.2 Identify Accuracy Verification Sources
- 4.3 Determine Applicable Data Correction Criteria for each Data Element
- 4.4 Determine Quality Standards (compliance levels)

5.0 Assess Current Level of Data Quality

- 5.1 Measure Data Quality
- 5.2 Validate and Define Data Definitions
- 5.3 Establish Statistical Controls

6.0 Measure Non-Quality Data Costs

- 6.1 Understand Business Performance Measures

6.2 Calculate Data Costs

6.3 Calculate Non-Quality Data Costs

7.0 Interpret and Report Data Quality

A.3 SAMPLE DATA QUALITY IMPROVEMENT PROCESS IMPROVEMENT WORK BREAKDOWN STRUCTURE

1.0 Plan Data Quality Improvement Process Project

2.0 Select Candidate Business Processes for Data Quality Improvement

2.1Conduct Root-Cause Analysis

2.2 Define Process Improvement(s)

3.0 Develop Plan for Data Quality Process Improvement

4.0 Implement Data Quality Improvement

5.0 Evaluate Impact of Data Quality Improvement

6.0 Standardize Data Quality Improvement

A.4 SAMPLE DATA QUALITY IMPROVEMENT PROCESS CORRECTION WORK BREAKDOWN STRUCTURE

1.0 Plan Data Quality Improvement Process Correction Project

2.0 Conduct Correction

2.1 Plan Data Correction

2.1.1 Refine Correction Approach, Plan, and Schedule

2.1.2 Identify and Prioritize Data to Be Corrected

2.1.3 Identify Method for Data Correction

2.2 Extract and Analyze Source Data

2.2.1 Plan and Execute Data Extraction

2.2.2 Analyze Data

2.2.3 Document Findings

2.3 Execute Manual and Automated Data Correction

2.3.1 Standardize Data for Atomic-Level Format and Values

2.3.2 Correct and Complete Data

2.3.3 Match and Consolidate Data

2.3.4 Analyze Defect Types

2.3.5 Transform and Enhance Data

2.3.6 Calculate Derived and Summary Data

2.3.7 Summarize Data Correction Activities

2.4 Determine Adequacy of Corrections

2.4.1 Identify the Data Elements/Groups Corrected (Content)

2.4.3 Re-Assess Data Element Quality

2.4.4 Identify Best Correction Practices

2.4.5 Recommend Improvements to Correction Tasks

A.5 SAMPLE DATA QUALITY IMPROVEMENT PROCESS CERTIFICATION WORK BREAKDOWN STRUCTURE

1.0 Certify Information Quality Process

2.0 Certify Data Corrections

3.0 Define Sample Size and Resources

4.0 Conduct Certification

5.0 Interpret and Report Information Quality Certification

APPENDIX B. DATA QUALITY SOFTWARE TOOLS

Data quality tools provide automation and management support for solving data quality problems.[22] Effective use of data quality tools requires:

- Understanding the problem to be solved
- Understanding the kinds of technologies available and their general functionality
- Understanding the capabilities of the tools
- Understanding any limitations of the tools
- Selecting the right tools based on existing requirements
- Using the tools properly.

Sections B.1-B.5, below, discuss five categories of data tools for data quality improvement and data correction that may be applied within individual Data Quality Projects at DOI to support the four-stage process for data quality improvement discussed in this Guide. It is recommended that each Data Quality Project Team choose its own tools to support specific program business needs. It is recommended that a Data Quality Project Team always select only one tool to accomplish a single category of data quality improvement. All software tools selected for data quality improvement and data correction must be in accordance with DOI's Technology Reference Model (TRM).

A caveat for the use of automated correction tools is that some varying percentage of the data must be corrected and verified manually by looking at hard-copy, "official" documents or by comparing them to the real-world object or event. Also, automated tools cannot ensure "correctness" or "accuracy."

B.1 DATA QUALITY ANALYSIS TOOLS

Automated tools may be used to conduct audits of data against a formal set of business rules to discover inconsistencies within those rules. Reports can be generated that depict the number and types of errors found. Quality analysis and audit tools measure the state of conformance of a database or process to the defined business rule.

B.2 BUSINESS RULE DISCOVERY TOOLS

Business rule discovery tools may be used to analyze legacy system data files and databases in order to identify data relationships that affect the data. This analysis may identify quantitative (formula-based) or qualitative (relationship-based) conditions that affect the data and its successful migration and transformation. The analysis may also identify exceptions or errors in the conditions.

Business rule discovery tools use data mining or algorithms to analyze data to discover:

- Domain value counts.
- Frequency distributions of data values.
- Patterns of data values in non-atomic data, such as unformatted names and addresses or textual data.
- Formulas or calculation algorithms.
- Relationships, such as duplicate data within or across files.

- Similarities of items, such as spelling.
- Correlation of data values in different fields.
- Patterns of behavior that may indicate possible intentional or unintentional fraud.

It is important to remember that there may be performance problems when using these tools if the files are large or contain many fields. Performance problems may be minimized through random sampling or by making separate analytical runs against different sets of fields, grouped in ways that meaningful business rules are likely to emerge.

B.3 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 knowledge of in-house data and analysis to adequately use them. Data correction tools may be used to standardize data, identify data duplication, and transform data into correct sets 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.

B.4 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 the occurrence of problems. Defect prevention tools provide the same kind of functions as data correction tools. The difference is that they provide for discovery and correction of the errors during the online data creation process, rather than in batch mode.

B.5 METADATA MANAGEMENT AND QUALITY TOOLS

Metadata management and quality tools provide automated management and quality control of the development of data definitions and data architecture. The tools perform one or more of the following functions:

- Ensure conformance to data-naming standards.
- Validate abbreviations of the names of data.
- Ensure that the required components of data definition are provided.
- Maintain metadata for control of the data reengineering and correction processes.
- Evaluate data models for normalization.
- Evaluate database design for integrity, such as primary key to foreign key integrity, and performance optimization.

Metadata management and quality tools support the documentation of the specification of the data product. These tools cannot determine if data required for information consumers are missing, defined correctly, or even required in the first place. Data 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 information consumers.

B.6 EVALUATING DATA QUALITY TOOLS

Tool selection is second only to the business problem at hand in architecting a business solutions environment. The Data Quality Project Team should evaluate any software tool from the standpoint of how well it solves business problems and supports the accomplishment of the enterprise's business objectives and should try to avoid "vendor pressure" to buy tools *before* the requirements are developed.

Once the business problems are defined, the Data Quality Project Team should determine what category of data quality function automation is required. For example, the fact that a data warehouse is being developed does not automatically mean that the problem is correcting data for the warehouse. The real problem may be data defects at the source, and the business problem to be solved is that the data producers do not know who uses the data they create. Therefore, a data defect prevention tool is required to solve the real business problem.

All software tools purchased for data quality improvement and data correction must be in accordance with DOI's Technology Reference Model (TRM).

APPENDIX C. DATA QUALITY IMPROVEMENT PROCESS BACKGROUND

This section presents background information necessary to understand the evolution of thought in Total Quality Management, which is the foundation of DOI's Data Quality Improvement Process methodology of continuous data quality improvement. It notes the change caused by the shift in focus from an intrinsic definition of quality and the corresponding thinking that it cannot be managed to achieve total quality, to the focus on the customer and the achievement of total quality management.

EVOLUTION OF QUALITY – FROM INTRINSIC TO CUSTOMER CENTRIC

The manufacturing industry in the United States operated in a steady state from the end of World War II until the late 1970's, when it suffered a revolution caused by the 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 were 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; and third, the cost of disposal. In those cases in which 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 for defects in the finished products. He indicated that inspection should be minimized and used only to determine if product variability is unacceptable, and he 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, was also based on the "PDCA" approach (continuous process improvement) developed by W. Stewart[23] and the Total Quality Management approach developed by P. B. Crosby.[24] 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.[25]

APPENDIX D. THE "ACCEPTABLE QUALITY LEVEL" PARADIGM

Philip Crosby makes the business case for non-quality: "There is absolutely no reason for having errors or defects in any product or service."[26] "It is much less expensive to prevent errors than to rework, scrap, or service them," because the cost of waste can run as much as 15 to 25 percent of sales.[27]

Crosby further states:

> "Now what is the existing standard for quality?
>
> "Most people talk about an AQL—an acceptable quality level. An AQL really means a commitment before we start the job to produce imperfect material. Let me repeat, an *acceptable quality level is a commitment before we start the job to produce imperfect material*. An AQL, therefore, is not a management standard. It is a determination of the status quo. Instead of the managers setting the standard, the operation sets the standard….
>
> "The Zero Defects concept is based on the fact that mistakes are caused by two things: lack of knowledge and lack of attention.
>
> "Lack of knowledge can be measured and attacked by tried and true means. But lack of attention is a state of mind. It is an attitude problem that must be changed by the individual.
>
> "When presented with the challenge to do this, and the encouragement to attempt it, the individual will respond enthusiastically. Remember that Zero Defects is not a motivation method, it is a performance standard. And it is not just for production people, it is for everyone. Some of the biggest gains occur in the non-production areas."[28]

The same is true for data quality. Larry English's analysis concludes that the costs of non-quality data can be as much as 10 to 25 percent of operating budgets and can be even higher in data intensive organizations.[29] In the absence of a set data quality standard, the standard is simply: "If data are required for business processes, what is the business case for errors or omissions when creating it? There is absolutely no reason for errors or defects in any data you create if those data are needed for other processes."[30]

The approach to reach the appropriate level of quality, or quality standard, for an information group, is to establish a customer-supplier agreement. These agreements are tailored to the situation and to the specific needs of the customers of the data, both short and long term, and they are signed and monitored by both the providers and customers of the data. Over time, these agreements can be improved to drive out the costs of waste due to scrap and rework. However, before an agreement can be put into place, the producing processes must be in control; that is, they must have predictable results. If the processes that produce needed data are not in control, the first customer-supplier contract needs to include a "*Standardize*-Do-Check-Act" to define the processes and put them in control. Once the processes are in control and their results are predictable and known, the parties have the proper foundation to reach an agreement for the quality target in the next time period.

APPENDIX E. ADDITIONAL LEGISLATION/REGULATIONS INFLUENCING DOI'S GUIDE

Additional legislation and/or regulations that support DOI's Data Quality Management are:

- The Federal E-Gov Act Section 207d[31] states that agencies use of standards to enable government information in a manner that is electronically searchable and interoperable across agencies, as appropriate.

- OMB Memo 06-02-Improving Public Access to data and Dissemination of Government Information[32] uses the Data Reference Model (DRM) to organize and categorize agency information intended for public access and make it searchable across agencies.

- OMB Circular A-16-Coordination of Geographic Information and Related Spatial Data Activities[33] to identify proven practices for the use and application of agency data sets.

- Federal Register Vol.67, No. 36, Friday, February 22, 2002[34] (Federal Register notice of Public Law 106-554 section 515 requirements), which states that agencies shall adopt a basic standard of quality appropriate for the various categories of information they disseminate. Agencies shall treat information quality as integral to every step of an agency's development of information, including creation, collection, maintenance, and dissemination.

- DOI's Department Manual Part 378[35] – Data Resource Management Policy, which has three key points:

    1. Manage Data as a Department asset.

    2. Reuse existing standards before creating new ones.

    3. Establishes core roles and responsibilities for enterprise data management.

- DOI's CIO Directive for Data Standardization, which further elaborates the applicable policy roles and provides a method and process for standardizing, publishing, and implementing data standards.

- The Clinger-Cohen Act of 1996, Public Law 104-106[36] (formerly the Information Technology Management Reform Act of 1995)

  The Clinger-Cohen Act assigns overall responsibility for the acquisition and management of Information Technology (IT) in the federal government to the Director, Office of Management and Budget (OMB). It also gives authority to acquire IT resources to the head of each executive agency and makes them responsible for effectively managing their IT investments.

  Among other provisions, the act requires agencies to:

    - Base decisions about IT investments on quantitative and qualitative factors associated with the costs, benefits, and risks of those investments.

- Use performance data to demonstrate how well the IT expenditures support improvements to agency programs.
- Appoint CIOs to carry out the IT management provisions of the act and the broader information resources management requirements of the Paperwork Reduction Act.

The Clinger-Cohen Act also encourages agencies to evaluate and adopt best management and acquisition practices used by private and public sector organizations.

The focus of this act is on requiring agencies to develop and maintain an integrated information technology architecture. This architecture can help: (1) ensure that an agency invests only in integrated, enterprise-wide business solutions and (2) move resources away from non-value-added, legacy business systems and non-integrated system development efforts.

- The Paperwork Reduction Act (PRA) of 1995[37], as amended (44 United States Code (U.S.C.) 3501-3520), was enacted largely to relieve the public of the mounting information collection and reporting requirements of the federal government. It also promoted coordinated information management activities on a government-wide basis by the Director of the Office of Management and Budget and prescribed information management responsibilities for the executive agencies. The management focus of the PRA was sharpened with the 1986 amendments that refined the concept of "information resources management" (IRM), defined as "the planning, budgeting, organizing, directing, training, promoting, controlling, and management activities associated with the burden, collection, creation, use, and dissemination of information by agencies, and includes the management of information and related resources."
- The Computer Matching and Personal Privacy Act of 1988[38] (P.L. 100-503), as amended, the Privacy Act of 1974, as amended (5 U.S.C. 552a (1995 and Supp. IV 1998), which states no agency shall disclose any record which is contained in a system of records by any means of communication to any person, or to another agency, except pursuant to a written request by, or with the prior written consent of, the individual to whom the record pertains.
- Freedom of Information Act (FOIA) of 1966[39], as amended (5 U.S.C. 522) in 2002, allows for the full or partial disclosure of previously unreleased information and documents controlled by the U.S. Government. The Act defines agency records subject to disclosure and outlines the mandatory disclosure procedures and grants nine exemptions to the statute.
- OMB Circular A-119[40], or the Voluntary Consensus Standards Policy, establishes policies on federal use and development of voluntary consensus standards and on conformity assessment activities. Pub. L. 104-113, the "National Technology Transfer and Advancement Act of 1995," codified existing policies in A-119, established reporting requirements, and authorized the National Institute of Standards and Technology to coordinate conformity assessment activities of the agencies. OMB is issuing this revision of the Circular in order to make the terminology of the Circular consistent with the National Technology Transfer and Advancement Act of 1995, to issue guidance to the agencies on making their reports to OMB, to direct the Secretary of Commerce to issue policy guidance for conformity assessment, and to make changes for clarity.

- DOI's Information Quality Guidelines implement the OMB Information Quality Guidelines. http://www.doi.gov/ocio/iq. These Guidelines address the process for ensuring the quality of information prior to dissemination. DOI's Data Quality Management Guide provides a more focused set of processes for monitoring and correcting data in DOI-owned data sources prior to dissemination as an information product. Examples of information disseminated products include web sites, reports, and manuals.

APPENDIX F. GLOSSARY

3-sigma (3σ or 3s): Three standard deviations used to describe a level of quality in which three 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 6.681%, allowing no more than 66,810 defects per million parts.

4-sigma (4σ or 4s): Four standard deviations used to describe a level of quality in which four 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 .621%, allowing no more than 6,210 defects per million parts**.**

6-sigma (6σ 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 the information consumer or end customer is able to access or get the data he or she needs (a component of Information Quality).

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

Accuracy to surrogate source: A measure of the degree to which data agree with an original, acknowledged authoritative source of data about a real-world object or event, such as a form, document, or unaltered electronic data received from outside the organization.

Atomic data values: A data value that is complete in and of itself, without reference to other data values. Numbers, strings, and ODB addresses are examples of atomic-data values.

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 data, i.e., "First Name" and "Last Name."

Authoritative Data Source (ADS): A single, officially-designated source authorized to provide a type or many types of information that is trusted, timely, and secure on which lines of business rely.

Automated data quality assessment: Data quality inspection using software tools to analyze data for business rule conformance. Automated tools can assess that a data element content is valid (adheres to business rules) for most business rules, and they can determine consistency across files or databases, referential integrity, and other mechanical aspects of data quality. However, they may not automate assessment of some very complex business rules, and they *cannot* evaluate accuracy.

Business concept: A person, place, thing, event, or idea that is relevant to the business and for which the enterprise collects, stores, and applies data. Procedural note: for business concepts to be properly used and managed, they must be clearly understood; this requires that they be concisely defined, using rigorous, declarative language (as opposed to procedural language).

Business data steward: The person who manages or the 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.

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

CASE: Acronym for *Computer-Aided Systems (or Software) Engineering*. The application of automated technologies to business and data modeling and systems (or software) engineering.

Common term: A Standard English word used by DOI as defined in a commercial dictionary (for instance "Enterprise is a unit of economic organization or activity, esp.: a business organization").

Completeness: A data quality dimension measuring the degree to which the required data are known. (1) *Fact* completeness is a measure of data definition quality expressed as a percentage of the attributes about an entity type that must be known to ensure that they are defined in the model and implemented in a database. For example, "80 percent of the attributes required to be known about customers have fields in a database to store the attribute values." (2) *Value* completeness is the first measure of data content quality expressed as a percentage of the required columns or fields of a table or file that actually have values in them. For example, "95 percent of the columns for the customer's table have a value in them." Value completeness is also referred to as *Coverage.* (3) *Occurrence* completeness is the second measure of the data content quality expressed as a percentage of the rows or records of a table or file that should be present in them. For example, "95 percent of the households which DOI needs to know about have a record (row) in the household table."

Concurrency: A data quality dimension measuring the degree to which the timing of equivalence of data is stored in redundant or distributed database files. The measure data concurrency may describe the minimum, maximum, and average data float time from when data are available in one data source and when they become available in another data source; or it may consist of the relative percent of data from a data source that is propagated to the target within a specified time frame. (Also, see *Data float.*)

Consistency: A data quality dimension 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 the information consumer or end customer to understand the meaning of the data and avoid misinterpretation (a component of Information Quality).

Controllable mission-critical data: Controllable means that DOI has control over data or data content because it is collected following DOI's standards or produced within DOI. Non-controllable mission-critical data are data acquired by DOI from sources that cannot be controlled by DOI, such as survey data from an external source or data resulting from DOI's actions, such as surveys from small samples with large margins of error, non-respondents who impact the representativeness of the sample, or inaccurate responses from respondents.

Data: The representation of facts. Data can represent facts in many media or forms including digital, textual, numerical, or graphical. The raw material from which information is produced when it is put in context that gives it meaning. (1) *Raw data* are data units that are used as the raw material in a defined 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 architecture: A "blueprint" of an enterprise expressed in terms of a business process model, showing what the enterprise does; an enterprise data model, showing what data resources are required; and a business data model, showing the relationships of the processes and data.

Data architecture quality: The degree to which data architecture correctly represents the structure and requirements of the data needs for a business area or process.

Data correction: See *Data Product Improvement*.

Data content quality: The subset of data quality referring to the quality of data values.

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

Data definition quality: The degree to which data definition accurately, completely, and understandably defines the meaning of the data being described.

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

Data dissemination: see *Dissemination of data*.

Data element: The smallest unit of named data that has meaning to an information consumer. A data element is the implementation of an attribute. Synonymous with data item and *field*.

Data improvement: See *Data Product Improvement*.

Data intermediary: a role in which individuals transform data from one form, not created by them, to another form (e.g., data entry technicians).

Data quality (assessed level): The measurement of actual quality of a set of data against its required quality dimensions.

Data quality (desired level): The level of data quality required to support the business needs of all data consumers.

Data quality assessment: The random sampling of a collection of data and testing the sample against their valid data values to determine their accuracy and reliability. Also called *data audit*.

Data quality: Data are of high quality if they are fit for their intended uses in operations, decision making and planning (J.M. Juran). Data relevant to their intended uses and of sufficient detail and quantity, with a high degree of accuracy and completeness, consistent with other sources, and presented in appropriate ways.

Data quality classes: There are three classes or tiers of data quality; Absolute Tier (Near Zero-Defect Data). Indicates these data can cause significant process failure when containing defects. Second Tier (High-Quality Data). Indicates there are high costs associated with defects in these data, and, therefore, it is critical to keep defects to a minimum. Third Tier (Medium-Quality Data). Indicates the costs associated with defects in these data are moderate and should be avoided whenever possible. When there is no impact associated with data defects, it may be an indication that the Department does need the data at all.

Data reengineering: The process of analyzing, standardizing, and transforming data from un-architected or non-standardized files or databases into enterprise-standardized data architecture (definition and architecture).

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

Data standardization: See *Data Definition*.

Data steward: There are seven business roles in data stewardship and nine information systems roles in data stewardship. See *Business data steward*.

Data validity source: The source of the data that provides the basis for determining the data entered into the database are valid or correct.

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

Definition conformance: the degree of consistency of the meaning of the actual data values with its data definition.

Dependency rules: The restrictions and requirements imposed upon the valid data values of a data element by the data value of another data element. Dependency rules are revealed in the business rules. Examples of dependency rules include:

- An Order without a Customer Name is not valid.
- If Employee Marital Status is 'Married.'
- Employee Spouse data must be present.
- An Employee Termination Date is not valid for an Active Employee.
- When an Order is 'Shipped,' the Order Shipping Date must be indicated.

Derivation integrity: a data quality dimension measuring the correctness with which derived data are calculated from their base data.

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

Dissemination of information: (In the context of information dissemination by federal agencies) Agency initiated or sponsored distribution of information to the public (see 5 CFR 1320.3(d) (definition of ``Conduct or Sponsor"')). Dissemination does not include distribution limited to government employees or agency contractors or grantees; intra- or inter-agency use or sharing of government data; and responses to requests for agency records under the Freedom of Data Act, the Privacy Act, the Federal Advisory Committee Act, or other relevant laws. This definition also does not include distribution limited to correspondence with individuals or persons, press releases, archival records, public filings, subpoenas, or adjudicative processes.

Dissemination: to spread abroad as if sowing seed (to plant seed for growth especially by scattering; e.g., disseminating ideas); to disperse throughout.

(DOI's) Data Quality Improvement Process: Techniques, methods, and management principles that provide for continuous improvement of the data processes of an enterprise. A management approach used by DOI, based upon accepted industry standards and incorporating project management and total quality management principles.

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.

Enterprise: a unit of economic organization or activity; especially: a business or government organization.

Extract, Correction, Transformation, and Load (ECTL): The process that extracts, corrects (or cleans), and transforms data from one database and loads it into another database, normally a data warehouse for an enterprise.

External partner: These are individuals and organizations that provide to and/or receive from DOI services and/or data regarding the Department. They include state and local governments, other federal agencies, and public service organizations.

Fact: The quality of being actual; something that has actual existence; an actual occurrence; a deed.

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.

Hidden data: Data stored within a defined data element that do not match the data element's definition.

Influential data: Scientific, financial, or statistical data from which a U. S. Government Agency can reasonably determine that dissemination of the data will have or does have a clear and substantial impact on important public policies or important private sector decisions.

Information (1): the communication or reception of knowledge or intelligence; knowledge obtained from investigation, study, or instruction; intelligence; news; facts, data; the attribute inherent in and communicated by one of two or more alternative sequences or arrangements of something (as nucleotides in DNA or binary digits in a computer program) that produce specific effects; a signal or character (as in a communication system or computer) representing data; something (as a message, experimental data, or a picture) that justifies change in a construct (as a plan or theory) that represents physical or mental experience or another construct; a quantitative measure of the content of data –*specifically,* a numerical quantity that measures the uncertainty of the outcome of an experiment to be performed.

Information (2): (In the context of business and government use; disseminated or not; this is the definition used in this Guide.) Data in context. The meaning given to data or the interpretation of data based on its context. It is the finished product as a result of the interpretation of data.

Information (3): (In the context of data dissemination by federal agencies) 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 data that an agency ***disseminates*** from a web page, but does not include the provision of hyperlinks to data 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 consumer: The role of individuals in which they use data in any form as part of their job function or in the course of performing a process, whether operational or strategic. Also referred to as a *data consumer* or *customer*. Accountable for work results created as a result of the use of data and for adhering to any policies governing the security, privacy, and confidentiality of the data used. The term information consumer was created by and has been used consistently by Peter Drucker since as early as 1973 to describe in general all "workers" in the Data Age organization.

Information float: The length of the delay in the time a fact becomes known in an organization to the time at which an interested information consumer 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 of 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 information consumer.

Information group: A relatively small and cohesive collection of data, consisting of 20–50 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 who use the data and several business roles that use the data.

Information presentation quality: Measuring the degree to which information-bearing mechanisms, such as screens, reports, and other communication media are easy to understand, efficient to use, and minimize the possibility of mistakes in their use.

Information producer: The role of individuals in which they originate, capture, create, or maintain data or knowledge as a part of their job functions or as part of the processes they perform. Information producers create the actual data content and are accountable for their accuracy and completeness to meet all data stakeholders' needs. See also *Data intermediary*.

Information product improvement: The process of data correction, reengineering, and transformation required to improve existing defective data up to an acceptable level of quality. This can be achieved through manual correction (by inspection or verification), manual or automated completion, filtering, merging, decoding, and translating. This is one component of *data scrap and rework*. See also *Data reengineering*. Information product improvement is *reactive* data quality.

Information quality: (1) The degree to which information consistently meets the requirements and expectations of the information consumers in performing their jobs. (2) *Assessed Information Quality:* The measurement of actual quality of a set of information against its required quality characteristics.

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 their origination to their 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 data are not compromised through corruption or falsification.

Logical data model: An abstract, formal representation of the categories of data and their relationships in the form of a diagram, such as an entity-relationship diagram. A logical data model is process independent, which means that it is fully normalized and, therefore, does not represent a process dependent (e.g., access-path) database schema.

Metadata: A term used to mean data that describe or specify other data. The term *metadata* is used to define all of the dimensions that need to be known about data in order to build databases and applications and to support information consumers and data producers.

Mission-critical data: Are data that are considered fundamental for DOI to conduct business or data frequently used by the Department, particularly financial data, key to the Department's integrity and accountability, and data used to support Government Performance and Results Act (GPRA) reports, Data Quality Projects teams, the Office of the Chief Financial Officer (OCFO), and the Office of the Chief Data Officer (OCIO). The Deputy Secretary or the Secretary will identify the data that will be categorized as mission-critical. Mission-critical data will be managed using the DATA QUALITY IMPROVEMENT PROCESS approach to enable DOI to achieve expected levels of data quality necessary to serve its constituents properly. (Also, see controllable, mission-critical data.)

Non-atomic data values: A data value that consists of multiple data values and which is logically complete only if all of its constituent values are defined. Non-atomic data values can temporarily take on invalid states while being updated, as multiple constituent parts are individually written.

Non-duplication: A data quality dimension that measures the degree to which there are no redundant occurrences of data.

Objectivity: The state whereby disseminated information is being presented in an accurate, clear, complete, and unbiased manner. This involves whether the information is presented within a proper context. Sometimes, in disseminating certain types of information to the public, other information must also be disseminated in order to ensure an accurate, clear, complete, and unbiased presentation. Also, the

agency needs to identify the sources of the disseminated information (to the extent possible, consistent with confidentiality protections) and, in scientific, financial, or statistical contexts, the supporting information and models so that the public can assess for itself whether there may be some reason to question the objectivity of the sources. Where appropriate, information should have full, accurate, transparent documentation, and error sources affecting data quality should be identified and disclosed to data consumers.

Physical Data Quality Assessment: Physical assessments compare data values to the real-world objects and events that the data represent in order to confirm that the values are accurate. This type of testing is more time and labor intensive than automated testing, but it is a necessity for confirming the accuracy of data. Physical assessments are usually complementary and must be consistent with and complementary to the corresponding automated assessment.

Plan, Do, Check, and Act (PDCA): An iterative, four-step quality control strategy. It is also referred to as the Shewhart cycle and was made popular by Dr. W. Edwards Deming in his Six Sigma programs. PLAN establishes the objectives and processes necessary to deliver results in accordance with the specifications. The cycle is executed as follows: DO implements the processes. CHECK monitors and evaluates the processes and results against objectives and Specifications and reports the outcome. ACT applies actions to the outcome for necessary improvement. This means reviewing all steps (Plan, Do, Check, Act) and modifying the process to improve it before its next implementation.

Precision: A data quality dimension measuring 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 square feet).

Primary key uniqueness: The prerequisite of a primary key to identify a single entity, row in a database, or occurrence in a file.

Primary key: The attributes that are used to uniquely identify a specific occurrence of an entity, relation, or file. A primary key that consists of more than one attribute is called a *composite* (or *concatenated*) primary key.

Process owner: The person responsible for the process definition and/or process execution. The process owner is the managerial data steward for the data created or updated by the process and is accountable for process performance integrity and the quality of data produced.

Quality standard: A mandated or required quality goal, reliability level, or quality model to be met and maintained.

Ranges, reasonability tests: General tests applied to data to determine if their values are correct. For example:

- A test for Birth Date on a Drivers License Application might be that the resulting age of the applicant be between 16 and 120.
- A range for a Patient's Temperature might be 80-110 ^{o}F, while the range for Room Temperature might be –20 to 120 ^{o}F.

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 readable source of operational data for these attributes or fields. In a fragmented data environment, a single occurrence may have different collections of fields that have records of reference 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: A data quality dimension measuring the degree to which related data conform to the associative business rules.

Repository: A database for storing data 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, and definitions. The repository acts as a basis for documentation and code generation specifications that will be used further in the systems development life cycle. Also referred to as *design dictionary*, *encyclopedia*, *object-oriented dictionary*, and *knowledge base*.

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.

Run Chart: Is a graph that displays observed data in a time sequence. Often, the information displayed represents some aspect of the output or performance of a manufacturing or other business process.

Scalability: The ability to scale to support larger or smaller volumes of data and more or less information consumers. The ability to increase or decrease size or capability in cost-effective increments with minimal impact on the unit cost of business and the procurement of additional services.

Surrogate source: a document, form, application, or other paper copy of the data from which the data were originally entered. Also, an electronic copy of the data generated outside the organization that are known to be accurate.

System of Records (SOR). As stated in the Privacy Act of 1974, a system of records is "a group of any records under the control of any agency from which information is retrieved by the name of the individual or by some identifying number, symbol, or other identifying particular assigned to the individual."
System stakeholder: One that has a stake or an interest or share in a system.

Synchronized secondary stores: Data that are coordinated copies of other, original data.

Timeliness: A data quality dimension measuring the degree to which data are available when information consumers or processes require them.

Unsynchronized secondary stores: Data that are copies of other, original data that are not coordinated with any action on the original data.

Usability: The degree to which the information presentation is directly and efficiently applicable for its purpose (a component of Information Quality).

User: A term used by many to refer to the role of people in data technology, computer systems, or data. The term is inappropriate to describe the roles of information producers and information consumers who perform the value work of the enterprise, the roles of those for whom information technology should enable them to transform their work, and the roles of those who depend on data to perform their work. With respect to information technology, applications, and data, the roles of business personnel are to produce and consume information. The term information consumer was created by and has been used consistently by Peter Drucker since 1973 to describe in general all "workers" in the Information-Age organization. The relationship of business personnel to information systems personnel is not as "users," but as *partners* who work together to solve the data and other problems of the enterprise.

Utility: The usefulness of the information to its intended consumers, including the public (a component of Information Quality). In assessing the usefulness of information that the agency disseminates to the

public, the agency must consider the uses of the information from the perspective of the agency and 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 data quality dimension measuring the degree to which the data conform 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.

Value and Cost Chain Diagram: Diagram that documents the processes that gather and hold a logical group of data from knowledge origination to the final database.

Work Breakdown Structures (WBS): A document that defines the work to be done and the personnel assigned to individual tasks as part of planning a Data Quality Assessment.

Zero defects: A state of quality characterized by defect-free products or 6-Sigma level quality. See *6 Sigma*.

APPENDIX G. FOOTNOTES

[1] http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=105_cong_public_laws&docid=f:publ227.105
[2] http://elips.doi.gov/elips/DM_word/3713.doc
[3] Glossary of Terms: Federal Enterprise Architecture Data Reference Model, Version 2.0 (November 17, 2005)
[4] Juran, Joseph M. and A. Blanton Godfrey, *Juran's Quality Handbook*, Fifth Edition, p. 2.2, McGraw-Hill, 1999
[5] http://www.doi.gov/ocio/architecture/documents/DOI%20Data%20Standardization%20Procedures%20-%20April%202006.doc
[6] English, Larry, *Improving Data Warehouse & Business Information Quality*, p. 123-124.
[7] English, Larry, *Improving Data Warehouse & Business Information Quality*, p. 123.
[8] English, Larry, *Improving Data Warehouse & Business Information Quality*, p. 83-118 for additional discussion of the critical quality characteristics; 119-123 for additional explanation of the procedures and techniques for this task.
[9] English, Larry, *Improving Data Warehouse & Business Information Quality*, p. 171.
[10] English, Larry, *Improving Data Warehouse & Business Information Quality*, p.167-188.
[11] English, Larry, *Improving Data Warehouse & Business Information Quality*, p. 188-196.
[12] English, Larry, *Improving Data Warehouse & Business Information Quality*, p. 289-302
[13] English, Larry, *Improving Data Warehouse &Business Information Quality*, p. 290.
[14] Shewhart, W., *Statistical Method from the Viewpoint of Quality Control*: New York: Dover Publications, 1986.
[15] English, Larry, *Improving Data Warehouse & Business Information Quality*, p. 302-309.
[16] English, Larry, *Improving Data Warehouse & Business Information Quality*, p. 298-299.
[17] English, Larry, *Improving Data Warehouse and Business Data Quality*, p. 301-302
[18] English, Larry, *Improving Data Warehouse & Business Information Quality*, p. 237-283.
[19] English, Larry, *Improving Data Warehouse &Business Information Quality*, p. 252-257.
[20] *The ABCs of Data Quality* seminar; Brentwood, TN; Data Impact International, p. 36-37
[21] English, Larry, *Improving Data Warehouse & Business Information Quality*, p. 188-302; adapted for DOI.
[22] English, Larry, *Improving Data Warehouse and Business Data Quality*, Chapter 10
[23] Shewhart, W., *Statistical Method from the Viewpoint of Quality Control*: New York: Dover Publications, 1986.
[24] Crosby, P.B, *Quality is Free: The Art of Making Quality Certain*: New York: Penguin Group, 1979.
[25] Imai, M., "*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.
[26] Crosby, P.B, *Quality is Free: The Art of Making Quality Certain, op. cit*., p. 58.
[27] Crosby, P.B, *Quality is Free: The Art of Making Quality Certain,* p. 149.
[28] Crosby, P.B, *Quality is Free: The Art of Making Quality Certain,* p. 146-147.
[29] English, Larry, *Improving Data Warehouse & Business Information Quality*, p. 12.
[30] English, Larry, *Data Quality Management: What Managers Must Know and Do* seminar. Brentwood, TN: Data Impact International, 1998-2002, p. 1.2
[31] http://www.xml.gov/documents/completed/eGovAct.htm
[32] http://www.whitehouse.gov/omb/memoranda/fy2006/m06-02.pdf
[33] http://www.whitehouse.gov/omb/circulars/a016/a016_rev.html
[34] http://www.whitehouse.gov/omb/fedreg/reproducible2.pdf.
[35] http://elips.doi.gov/app_DM/act_getfiles.cfm?relnum=3713
[36] http://www.cio.gov/Documents/it_management_reform_act_Feb_1996.html
[37] http://www.archives.gov/federal-register/laws/paperwork-reduction/
[38] http://www.whitehouse.gov/omb/inforeg/final_guidance_pl100-503.pdf

[39] http://www.usdoj.gov/oip/foiastat.htm
[40] http://www.whitehouse.gov/omb/circulars/a119/a119.html