Department of Housing and Urban Development

Total Information Quality Management Handbook

3300.1

Prepared by the Office of the Chief Information Officer

May 2003

FOREWORD

May 2003

This document is being issued as a new handbook to the "Common Data Element Correction Method: A Guidebook, Version 1," first distributed in May 1998 and revised in October 1999. Four principal drivers are behind the development of this handbook:

- Office of the Inspector General audit recommendations in March 2001 (OIG 2001-FO-0004) stressed the need for continuous quality improvement in HUD's mission-critical information. The OIG report recommended that the Office of the Chief Information Officer (OCIO): "(1) Implement data quality standards for systems and data supporting performance indicators, (2) Require data quality plans for these systems based on a standard rationale, and (3) Implement a methodology for independent verification for high priority data." This Handbook provides guidance for continuous information quality improvement activities within the Department.
- This Handbook brings HUD information quality practice in compliance with Section 515 of the Treasury and General Government Appropriations Act for Fiscal Year 2001 (Public Law 106-554; 114 Stat. 2763, hereafter referred to as "Section 515") and with the Office of Management and Budget (OMB) issued government-wide guidance that "provide policy and procedural guidance to Federal agencies for ensuring and maximizing the quality, objectivity, utility and integrity of information … disseminated by Federal Agencies." The "HUD Final Information Quality Guidelines" were published in the Federal Register on November 18, 2002 and are available at http://www.hud.gov:80/offices/adm/grants/qualityinfo/fr4769n02-final.pdf.
- This Handbook adds sections on data correction methods and includes guidance required to implement a Total Information Quality Management (TIQM®) approach at HUD. In addition, Chapter 5 is added to incorporate and update the "FSI Independent Verification Method," first released by the Financial Systems Integration (FSI) Data Quality Team in August 1999.
- The Handbook incorporates "lessons learned" from HUD's FY 2001 data correction efforts.

Send recommended changes to the document to:

OCIO Data Architect
470 East L'Enfant Plaza, Room 3113
Washington, D.C. 20410

Table of Contents

INTRODUCTION iii

CHAPTER 1. IMPLEMENTING HUD'S TOTAL INFORMATION QUALITY MANAGEMENT ENVIRONMENT 1-1

1-1 Overview 1-1
1-2 Definition of Information 1-1
1-3 Definition of Information Quality 1-1
1-4 HUD Mission-Critical Information Quality Standards 1-2
1-5 HUD TIQM® Roles and Responsibilities 1-4
1-6 HUD TIQM® Overview 1-5

CHAPTER 2. TIQM® ASSESSMENT PROCESS 2-1

2-1 Overview 2-1
2-2 Select Information Group Candidates 2-1
2-3 Assess Data Definition and Information Architecture Quality 2-6
2-4 Determine Desired Quality Standards for Prioritized Data Elements 2-9
2-5 Assess Current Level of Information Quality 2-12
2-6 Measure Non-Quality Information Costs 2-13
2-7 Interpret and Report Information Quality State 2-14

CHAPTER 3. TIQM® IMPROVEMENT PROCESS 3-1

3-1 Overview 3-1
3-2 Select Process for Information Quality Improvement 3-1
3-3 Develop **Plan** for Information Quality Process Improvement 3-2
3-4 **Do** Implement Information Quality Improvement 3-5
3-5 **Check** Impact of Information Quality Improvement 3-5
3-6 **Act** to Standardize Information Quality Improvement 3-5

CHAPTER 4. TIQM® CORRECTION PROCESS 4-1

4-1 Overview 4-1
4-2 Plan Data Correction 4-1
4-3 Extract and Analyze Source Data 4-2
4-4 Execute Manual and Automated Data Correction 4-5
4-5 Determine Adequacy of Correction 4-8

CHAPTER 5. TIQM® CERTIFICATION PROCESS 5-1

5-1 Overview 5-1
5-2 Certify Information Quality Process Improvements 5-1
5-3 Certify Data Corrections 5-2

List of Appendices

APPENDIX A. TIQM® PLANNING A-1

A.1 OUTLINE FOR TIQM® A-1
A.2 SAMPLE TIQM® ASSESSMENT WORK BREAKDOWN STRUCTURE A-1
A.3 SAMPLE TIQM® IMPROVEMENT WORK BREAKDOWN STRUCTURE A-2
A.4 SAMPLE TIQM® CORRECTION WORK BREAKDOWN STRUCTURE A-3
A.5 SAMPLE TIQM® CERTIFICATION WORK BREAKDOWN STRUCTURE A-3

APPENDIX B. INFORMATION QUALITY SOFTWARE TOOLS B-1

B.1 INFORMATION QUALITY ANALYSIS TOOLS B-1
B.2 BUSINESS RULE DISCOVERY TOOLS B-1
B.3 DATA REENGINEERING AND CORRECTION TOOLS B-2
B.4 DEFECT PREVENTION TOOLS B-2
B.5 METADATA MANAGEMENT AND QUALITY TOOLS B-2
B.6 EVALUATING INFORMATION QUALITY TOOLS B-3

APPENDIX C. TOTAL INFORMATION QUALITY MANAGEMENT CONCEPTS C-1

C.1 EVOLUTION OF QUALITY – FROM INTRINSIC TO CUSTOMER CENTRIC C-1
C.2 THE "ACCEPTABLE QUALITY LEVEL" PARADIGM C-1

APPENDIX D. GLOSSARY D-1

APPENDIX E. FOOTNOTES E-1

Table of Figures

Figure 1.1: HUD Mission-Critical Information Quality Standards 1-3

Figure 1.2: HUD's Information Quality Organization 1-5

Figure 1.3: HUD's Total Information Quality Management Method 1-7

Figure 2.1: Illustration of a Data Element Scope Worksheet 2-2

Figure 2.2: Characteristics of Information Content Quality 2-4

Figure 2.3: Characteristics of Information Presentation Quality 2-4

Figure 2.4: Information Quality Assessment Point by Assessment Objective 2-5

Figure 2.5: Illustration of a Data Element Prioritization Worksheet 2-6

Figure 2.6: Illustration of the Business Concept Template 2-8

Figure 2.7: Illustration of a Data Element By Record of Origin Worksheet 2-9

Figure 2.8: Illustration of Quality Target Compliance 2-11

Figure 2.9: Illustration of an Assessment Procedure Report Template 2-15

Figure 3.1: Illustration of a Cause and Effect Diagram 3-3

Figure 3.2: Cause and Effect Diagram Template for Information Quality 3-3

Figure 4.1: Illustration of a Data Definition Worksheet 4-3

Figure 4.2: Illustration of a Data Mapping Worksheet 4-4

Figure 4.3: Illustration of a Data Correction Worksheet Template 4-6

INTRODUCTION

It is the responsibility of the Department of Housing and Urban Development (HUD) to ensure that its financial and programmatic information have a level of quality that makes the information credible and useful for all its intended business purposes, within and beyond the Department. The Department's growing concern with the quality of this information, along with the Secretary's desire to report accurately where and how HUD dollars are being spent to revitalize communities across America, led the Secretary to charge the Chief Financial Officer (CFO) in 1998 with responsibility for data quality. As an initial effort, the CFO identified a dedicated team to lead the implementation of data quality principles. In Fiscal Year (FY) 1998, this team inaugurated a standard data quality cleanup method across HUD programs and organizations.

In FY 2000, the Office of the Chief Information Officer (OCIO) assumed responsibility for data quality. Working in partnership with the CFO and Program Areas and in response to Office of the Inspector General (OIG) audit recommendations,[1] the Chief Information Officer (CIO) began to evolve a systematic program for Total Information Quality Management for mission-critical information in HUD's major information systems. Mission-critical information is considered fundamental for HUD to conduct business, or information frequently used by the Department, particularly financial information, key to the Department's integrity and accountability, and information used to support Annual Performance Plans. The CIO and CFO together chartered a Data Control Board (DCB), responsible for steering the Department's information quality improvement practice.

Total Information Quality Management is a four-stage process: assessment, process improvement, correction, and certification. OCIO is responsible for independent assessment and certification of mission-critical information, particularly information used to support Annual Performance Plans. Program Area Managers and systems sponsors are responsible for both process improvement and data correction, and may perform their own assessment and certification activities.

This Handbook is a central part of the Department's strategy to guide its efforts for continuous process improvement in the acquisition, creation, maintenance, storage, and application of information. It complements HUD's Enterprise Data Management (EDM) Policy 3260.1. This policy institutes the Enterprise Data Management Practice (EDMP) that seeks to align data management priorities with HUD's mission and Program Area objectives and to streamline data management functions across the enterprise. The OCIO is implementing its enterprise data management program concurrent with and in alignment with the development of its Enterprise Architecture (EA).

The use of the TIQM® method provides the following advantages:

- Relies on government and industry-accepted practices of applying information quality criteria for a consistent level of quality.
- Facilitates Department-wide monitoring of improvement and correction activities.
- Facilitates schedule integration and coordination.
- Provides common milestones/products.
- Institutionalizes information quality.
- Provides the capability to share techniques, solutions, and resources throughout the Department.

This Handbook provides the reader with the concepts, step-by-step processes, illustration of and references to worksheets that will guide and assist with the improvement and correction processes.

The Handbook is divided into the following sections:

- Chapter 1: Implementing HUD's Total Information Quality Management Environment. Provides the definition of information, information quality, the quality standard, roles and responsibilities and an overview of the TIQM® method.

- Chapter 2: TIQM® Assessment Process. Describes the assessment process from selection to final report.
- Chapter 3: TIQM® Improvement Process. Describes the method and techniques that shall be used by the Program Areas to perform information quality improvement.
- Chapter 4: TIQM® Correction Process. Describes the method and techniques that shall be used by the Program Areas to perform data correction.
- Chapter 5: TIQM® Certification Process. Describes the method and techniques that shall be used by the OCIO to perform the final task of independent verification or certification.
- Appendix A: TIQM® Planning. Contains guidance and examples of Work Breakdown Structures to plan Information Quality Assessment, Information Quality Improvement, Data Correction, or Certification projects.
- Appendix B: Information Quality Software Tools. Contains a discussion of available software tools that can be used to automate information quality management and improvement.
- Appendix C: Total Information Quality Management Concepts. Provides a brief discussion of the critical concepts behind TQM and TIQM®.
- Appendix D: Glossary. Provides a glossary of terms used in this document that require special attention.

CHAPTER 1. IMPLEMENTING HUD'S TOTAL INFORMATION QUALITY MANAGEMENT ENVIRONMENT

1-1 Overview

A. This chapter presents a definition of information and information quality, HUD's information quality standard, roles and responsibilities, and an overview of HUD's Total Information Quality Management method. The implementation of this method must be in the context of a cultural transformation characterized by:

1. A value system: "we value our information customers."
2. A mindset of "excellence in all our work products including information as a product."
3. A habit of continuous improvement: "to eliminate the waste of process failure and information scrap and rework caused by defective, non-quality data."

1-2 Definition of Information

A. In this Handbook, information refers to data in context. Data is the representation of facts in all media or forms including digital, written, coded, textual, numerical, graphical, geo-spatial, audio and video. 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. This document provides a method for total quality management applicable to all information in all forms required by HUD.

B. Based on OMB Section 515 guidance, agencies are directed to develop management procedures for reviewing and substantiating the quality of information before it is disseminated. In Section 515 guidelines, the term information is used primarily in the context of dissemination of information and correction of disseminated information. Given that the method described in this Handbook applies to all kinds of information, disseminated or not, it can be applied to the improvement and correction of all information as needed. This method complements HUD's Section 515 guidelines as follows:

1. **Utility** – the information is usable, supported through the TIQM® quality characteristics of timeliness, concurrency, precision, accessibility, contextual clarity, usability, and rightness.
2. **Objectivity** – the information is being presented in an accurate, clear, complete, and unbiased manner, and as a matter of substance, is accurate, reliable, and unbiased; supported through the TIQM® quality characteristics of accuracy to reality, accuracy to surrogate source, precision, validity, completeness, relationship validity, consistency, concurrency, contextual clarity, usability, derivation integrity and rightness.
3. **Integrity –** the information is protected from corruption or falsification, and is not addressed in this document. For more information on information protection, refer to the HUD Handbook 1750.1 Rev – 4, CHG-3, "National Security Information".

1-3 Definition of Information Quality

A. During the 1990's, Larry English took the proven quality principles of Deming, Shewhart, Crosby and Imai (for a brief discussion of the evolution of quality management refer to Section C.1 below) and adapted them to information management with the same results. Information is a product "manufactured" by one or multiple processes (taking a

loan or a grant application) and consumed by other processes (reporting performance indicators) or customers (public housing authorities).

B. The processes of "manufacturing" (creating and storing, maintaining, propagating and delivering) information have multiple customers, and these customers have multiple needs and expectations that the information product must meet. Information customers can be internal or external. Internal customers are processes and people consuming information to make critical decisions, such as underwriting an application or securing funding for future programs, providing insight into the Department's performance, or servicing the public. External customers include the public, state and local governments, Congress, public service organizations, and the Executive Branch. Customer needs and expectations include documented requirements as well as unwritten expectations.

C. **Information quality means consistently meeting the information customer's expectations.**[2]

D. Improving information quality involves correcting defective data and implementing quality improvement procedures that ensure the expected levels of information quality are achieved and maintained.

E. Information quality has three components:[3]

1. **Data Definition and Information Architecture Quality**: Proper information definition accurately describes the meaning of the real world object or event that the data represents and meets the needs of all information customers to understand the data they use. Proper information architecture correctly represents the structure of the inherent and real relationships of information to represent real-world objects and events and is stable and flexible. Data definition and information architecture are the specification of the information product and must represent the views and needs of all the business areas, applications, and end customers of the information. The data definition and information architecture include the business definition, the domain or value set, and the business rules that govern the data. For a detailed description of information definition and information architecture characteristics, refer to Section 2-3.
2. **Data Content Quality**: Content quality cannot be measured without a quality definition. Data content quality is the degree to which data values accurately represent the characteristics of the real-world object or event and meet the needs of the information customers to perform their jobs effectively. For a detailed description of data content quality characteristics, refer to Figure 2.2.
3. **Data Presentation Quality:** Data presentation quality is the degree to which the information presented enables the knowledge worker or end customer to apply the information efficiently and effectively. Data presentation quality has several dimensions, including accessibility, contextual clarity, usability, and rightness. For a detailed description of information presentation quality characteristics, refer to Figure 2.3.

1-4 HUD Mission-Critical Information Quality Standards

A. HUD must set information quality standards based upon downstream knowledge worker and external stakeholder expectations and requirements for the data. In order to provide the Program Areas with specific and actionable direction, this section describes specific standards for mission-critical information quality, especially information used to support Annual Performance Plans. Each Program Area is encouraged to use this framework to establish information quality standards for other information deemed critical by the Program Area.

B. The three components of information quality must be managed based on the quality class of the information to achieve total information quality. The quality class indicates the degree of quality required for the particular information based on business need. Information quality can be of three classes:

1. Absolute (zero-defect or close to zero-defect) indicates this information can cause significant process failure when containing defects. This includes mission-critical information and may include influential information.
2. Second Tier (high cost of non-quality) indicates there are high costs associated with defects in this information and therefore it is critical to keep defects to a minimum.
3. Third Tier (moderate cost of non-quality) indicates the costs associated with defects in this information are moderate and must be avoided whenever possible. If for some information there is no impact associated with defects, this indicates that the Department may not require the information at all.

C. Figure 1.1 shows HUD's quality standards for mission-critical information, which is of quality class "A".

	Measure	Short Term (12 Month)		Long Term (5 Years)	
		Target	Confidence	Target	Confidence
Definition	% Complete	95%	95%	100%	99%
	Customer Satisfaction	95%	95%	100%	99%
	Known and Acceptable Definition & Structure Defects	95%	95%	100%	99%
Content	Data Correctness (maximum Errors Per Million parts)	4 σ (6,210 EPM)	95%	4 σ (6,210 EPM)	99%
	Processes Producing Mission-Critical Information in Statistical Control	100%	95%	100%	99%
	Elimination of Known Defect Production (New Defects)Through Information Quality Improvement	50%	N/A	50% (per year)	N/A
Presentation	Accessibility	95%	95%	100%	99%
	Contextual Clarity	95%	95%	100%	99%
	Usability	95%	95%	100%	99%
	Rightness	95%	95%	100%	99%

Figure 1.1: HUD Mission-Critical Information Quality Standards

D. The data content quality standard in Figure 1.1 applies to controllable mission-critical information. Controllable means that HUD has control over the content, because it is collected following HUD standards (e.g., housing authority filings) or produced within HUD. 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.

1-5 HUD TIQM® Roles and Responsibilities

A. HUD's Enterprise Data Management (EDM) Policy 3260.1 establishes the roles and responsibilities for

1. **The Data Control Board** (DCB) to direct and facilitate the continuous improvement of information quality at HUD.
2. **The Enterprise Data Management Group** (EDMG) within the OCIO to establish HUD's EDMP in close coordination with the Enterprise Architecture plan and the DCB, and to execute the assessment and certification of mission-critical information.

B. Program Area managers are accountable for the quality of the information produced by the processes and applications in their charge and for ensuring that it meets the needs of all their information customers. Program Area managers are also accountable for the information they collect from their external stakeholders and should work with them to ensure they apply the appropriate information quality management discipline to the information they produce and provide to HUD. It is their responsibility to[4]

1. Implement information policy and ensure compliance,
2. Develop plans consistent with information sharing to maximize reuse and value of information and minimize information costs,
3. Implement information quality design principles and standards that meet all internal and external HUD stakeholders' expectations,
4. Implement effective business rules and provide resources and training to information producers to ensure quality data capture,
5. Provide resources and training to staff in information process improvement (Plan-Do-Check-Act) and empowering staff to improve processes to accomplish quality standards,
6. Present information in reports and screens that minimize ambiguity and maximize the information's objectivity and utility for the information customers.

C. **Program Area Quality Improvement Teams** (PAQIT) and **Correction Teams** (PACT) are ad-hoc teams chartered by the Program Areas to conduct information quality improvement and/or data correction efforts. Each Program Area may charter one or more teams as required for a specific Program Area project. These teams should include information producers and subject matter experts who understand the intended use of the data across the enterprise from initial creation to final disposal. Individuals may be members of both teams. Team members should be aware of all known and accepted data values for a data element as well as the environment in which the data is captured. Under the discretion and direction of the Program Areas, these teams may also conduct internal assessments of Program Area critical information.

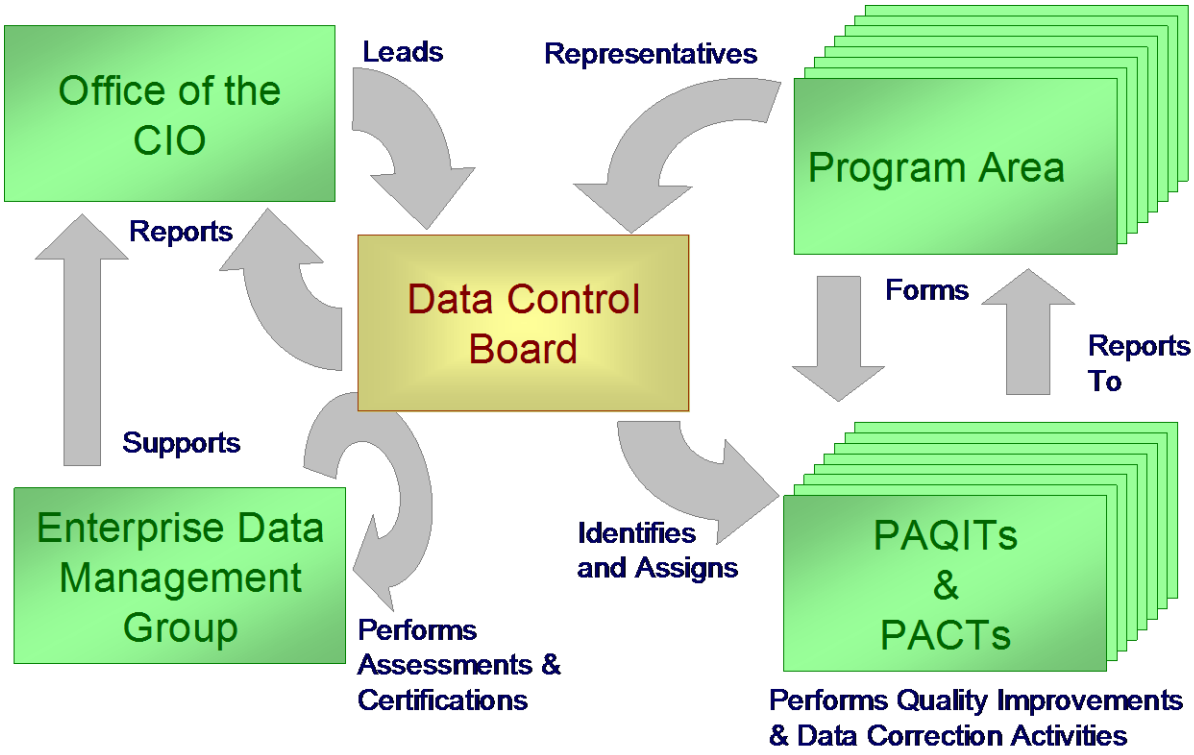

Figure 1.2: HUD's Information Quality Organization

1-6 HUD TIQM® Overview

A. To assist the Department with its strategy to implement Total Information Quality Management, this Handbook includes the steps to

1. Identify, prioritize, and assess areas of opportunity,
2. Determine the most effective approach to improve processes to ensure that defective data is no longer produced,
3. Correct the existing defective data,
4. Certify that the process and the information are in compliance with expected levels of quality or quality standards.

B. The TIQM® approach, shown in Figure 1.3, is based upon accepted industry standards and incorporates project management and total quality management principles. The method is iterative and may be repeated until the information reaches the appropriate quality levels. The following is a summary of each of the steps described in this Handbook; the numbers correspond to the chapters where the topic is presented in detail.

1. **Implementing HUD's Total Information Quality Management Environment**. Chapter 1 focuses on the systemic aspects that must be addressed within HUD to establish the proper environment for the successful deployment of a continuous information quality improvement.
2. **Assessment**. Chapter 2 focuses on the assessment of the state of information quality. The EDMG will execute this process in the assessment of mission-critical information. Program Areas 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 information architecture quality, determining the desired quality standards, assessing the current level of

information quality, measuring the non-quality information costs, and interpreting and reporting the state of information quality. The outcome of the assessment, as a part of the final report, is a set of recommended follow-on actions for OCIO and the Program Area with which both concur. The Data Control Board is accountable for reviewing and accepting final assessment reports.

3. **Improvement**. Once the assessment has identified areas of improvement, the Program Areas shall initiate activities to improve the quality of the information they acquire or produce. This is a proactive effort to prevent the incidence of defects in the data by attacking the causes of non-quality information. Improvement consists of selecting the process for information quality improvement, developing a plan for improvement, implementing the improvement in a controlled environment, checking of the impact of the improvement to make sure that results are as expected, and standardizing the improvement across the enterprise.
4. **Correction**. Once the assessment has identified areas of correction, the Program Areas shall initiate activities to correct the quality of the information they acquire or produce. 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.
5. **Certification**. As soon as Program Areas complete improvement and corrections efforts as described above, the EDMG independently verifies that the measures undertaken have accomplished the stated goals for mission-critical information. Certification consists of the certification of information quality improvements and/or the certification of data corrections. The objective is two–fold: first to ensure that mission-critical information is held to the agreed upon standards; and second, to ensure that the processes and procedures for improvement and correction are themselves improved over time for more effective and efficient results for HUD.

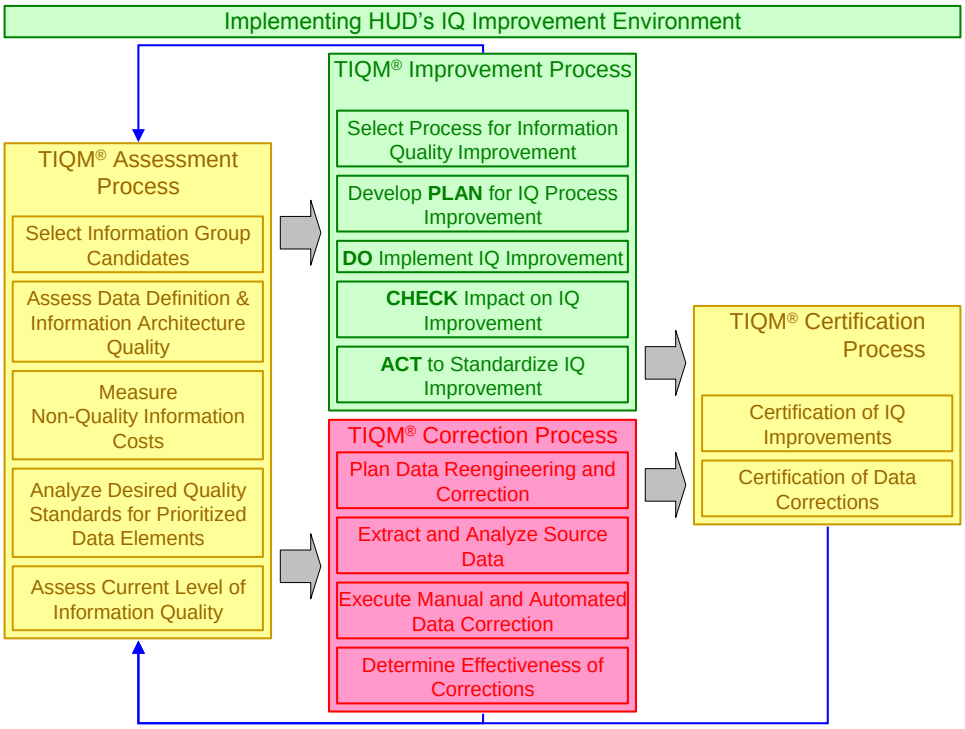

Figure 1.3: HUD's Total Information Quality Management Method

CHAPTER 2. TIQM® ASSESSMENT PROCESS

2-1 Overview

A. This chapter is concerned with the first stage of HUD's information quality improvement: assessment. Assessment is the first step to achieve expected levels of information quality necessary for HUD to serve its constituents properly.

2-2 Select Information Group Candidates

A. This step provides a guide for Program Areas to use when selecting 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 produces it. Therefore, a set of criteria for selecting and prioritizing which data elements to assess and improve must be developed. This is the process of determining the scope of a project.[5]

B. Determine Scope Based on Business Needs

1. Documenting the scope of effort is necessary to provide direction for information quality improvement and data correction. To obtain the most value from these efforts, it is necessary to assess business needs, taking into consideration the entire information value and cost chain that may be affected by information quality. To determine enterprise-wide business needs accurately, it will be necessary to conduct interviews with knowledge workers in each of HUD's business areas to find out how they and their information stakeholders outside of HUD are using the information and to determine their quality expectations.

C. Identify Information Group to be Assessed

1. Once the business needs have been defined, identify the data necessary to support those business needs. Collect information from knowledge worker interviews in each of HUD's business areas to determine how they use the information in the performance of their jobs, and how the data supports the business need. The objective is to determine information where assessment and improvement could yield significant tangible benefits.[6]

2. The following example in Figure 2.1 illustrates how to document the data necessary to support the business needs.

Data Element Scope Worksheet *Example*

Information Group	Data Element (Table or Record Name)	Within Scope? (Y/N)	Rationale for Inclusion or Exclusion (process and decision requiring it, and consequences if data is defective)
Inspection	Inspection Date	Y	• For selecting inspections within the last fiscal year. • Printed on the report.
	Inspection Report Completion Date	N	• Not applicable to this report or indicator.

Data Element Scope Worksheet *Example*			
Information Group	Data Element (Table or Record Name)	Within Scope? (Y/N)	Rationale for Inclusion or Exclusion (process and decision requiring it, and consequences if data is defective)
Human Resource	Inspector First Name, Middle Initial, Last Name	N	• Not applicable to this report or indicator.
Property	Property ID	Y	• To distinguish a particular property in the computer system.
	Property Name	Y	• Printed on the report.
	Property Street Address, City, State, Zip	Y	• Printed on the report.
	Property Contact Phone Number	N	• Not applicable to this report or indicator.
Organization	Regional HUD Office Street Address, City State, Zip	N	• Not applicable to this report or indicator.

Figure 2.1: Illustration of a Data Element Scope Worksheet

D. Identify Information Value and Cost Chain

1. For the Information Groups and Data Elements in the scope of the assessment, determine all business processes, applications, and people who create or update a group of data along with the dependencies between the processes that create or update the same data. This is to identify all points at which information quality can be impacted.

2. The information value and cost chain includes (source: *Improving Data Warehouse and Business Information Quality*, p. 160-162)

   a. All database and files, including paper documents, in which the data is stored from its point of origin to its last database of use;
   b. All business processes and application programs that create, update, or delete data;
   c. All replicate, extract and interface programs that copy data from one database and transform it and propagate it to another database;
   d. All information producer and knowledge worker roles in the process;
   e. All application programs that retrieve the data.

E. Identify Information Stakeholders

1. For the selected information group, identify the categories of information stakeholders. These stakeholders include

   a. The information producers (including Program Areas, support offices, other federal agencies, state and local governments, and lenders) that create or maintain the information;
   b. The knowledge workers who use it, including Program Areas and support offices within HUD;

c. The end customers, including Program Areas, support offices, the Executive Branch, Congress, GAO, OMB and taxpayers who require it.

For each of these categories, identify key contacts and maintain this information in an authoritative repository for future reuse.

F. Identify Information Quality Objectives and Measures

1. Establish the information quality characteristics to be measured in the information group to be assessed. There are two sets of characteristics: data content and presentation. The data content quality characteristics are described in Figure 2.2.

Characteristic	Quality Characteristic Description	Example of Non-Quality Data
Validity	The degree to which the data conforms to its definition, domain values and business rules.	A U.S. address has a state abbreviation that is not a valid abbreviation (not in the valid state abbreviation list).
Non-Duplication	The degree to which there are no redundant occurrences or records of the same real world object or event.	One applicant has multiple applicant records (evident when an applicant gets duplicate, even conflicting, notices).
Completeness	The degree to which all required data is known. This includes having all required data elements (all facts about the object or event), having all required records, and having all required values.	An indicator for spouse is set to “yes”, but spousal data is not present.
Relationship Validity	The degree to which related data conforms to the associative business rules.	A property address shows a Michigan zip code, but a Florida city and state.
Consistency	The degree to which redundant facts are equivalent across two or more databases in which the facts are maintained.	The same applicant is present in two databases or systems and has different name, address, or dependents.
Concurrency	The timing of updates to ensure that duplicate data stored in redundant files is equivalent. This is a measure of the information float (the time elapsed from the initial acquisition of the information in one file or table to the time it is propagated to another file or table).	On Monday, an applicant’s change of address is updated in the Applicant record of origin file, but the record is propagated to the main Program database after the weekend cycle (Friday night). That record has a concurrency float of 5 days between the record-of-origin file and the record-of-reference database.
Timeliness	The degree to which data is available to support a given knowledge worker or process when required.	A change of address is needed to schedule an inspection but is not available to the field office, and the inspector leaves without the proper information.
Accurate (to reality)	The degree to which data accurately reflects the real-world object or event being described.	The home telephone number for a customer record does not match the actual telephone number.

Characteristic	Quality Characteristic Description	Example of Non-Quality Data
Accurate (to surrogate source)	The degree to which the data matches the original source of data, such as a form, application, or other document	An applicant's reported income on the application form does not match what is in the database.
Precision	The degree to which data is known to the right level of detail (e.g., the right number of decimal digits to the right of the decimal point).	The summary amounts in congressional reports are rounded to the nearest $1,000.00 and do not include amounts in the hundreds, tens, dollars or pennies. However, the amounts will be aggregated in dollars and cents and then rounded to the nearest $1,000 to avoid rounding errors.
Derivation Integrity	The correctness with which derived data is calculated from its base data.	The summary of accounts for a given district does not contain all valid entries for the district.

(Source: *Improving Data Warehouse and Business Information Quality*, p. 142-143)

Figure 2.2: Characteristics of Information Content Quality

The presentation quality characteristics are listed in Figure 2.3.

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

(Source: *Improving Data Warehouse and Business Information Quality*, p. 142-143)

Figure 2.3: Characteristics of Information Presentation Quality

G. Determine Files and Processes to Assess

1. Depending upon the assessment objectives, information may need to be measured at different points in the information chain (see Figure 2.4 below). Identify the system(s) that capture, maintain or use the information group, and assess the same information in all applications and files or databases. There is a

tendency to assess the quality only in the circle of influence (the owned application or database); however, the critical impact to the Department occurs when the information is not of the expected quality and is shared across applications and Program Areas. If there are resource or time constraints, it is better to reduce the number of data elements in the assessment but include the entire value chain for the information being assessed. This means that the assessment must include all relevant databases, applications, files, and interfaces. In all cases, the approach to be taken must be defined and documented.

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 information.	A sample of data at the point of origin. These must be compared against equivalent data from the database from which timely access is required.
7. Ensure effectiveness of data warehouse conditioning process.	A sample of data from the record-of-reference. These must be compared against equivalent record(s) in the data warehouse.

(Source: *Improving Data Warehouse and Business Information Quality*, table 6.2, p. 165)

Figure 2.4: Information Quality Assessment Point by Assessment Objective

H. Prioritize Data Elements Supporting Business Need

1. Once the data elements necessary to support the business needs have been identified, prioritize the data elements. A simple high-medium-low scale may be used. Knowledge workers who understand how the data meets their requirements best make this determination.

2. As stated in Section 515 Guidelines, "The more important the information, the higher the quality standards to which it should be held." Factors making a data element high priority might be

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

   Figure 2.5 illustrates how to document the data element prioritization:

Data Element Prioritization Worksheet			
Information Group	Data Element (Table or Record Name)	Data Element Priority (High, Medium, Low)	Rationale for Priority (process and decision requiring it, and consequences if data is defective)
Inspection	Inspection Date	High	This is critical to determine if the inspection had been 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 information 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	Medium	This is an important characteristic of the property but is not indispensable for the report.
	Property Owner 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.

Figure 2.5: Illustration of a Data Element Prioritization Worksheet

2-3 Assess Data Definition and Information Architecture Quality

A. In this step, determine the quality measures for data definition and information architecture. Also, evaluate the structure and definition of the information under assessment. Finally, develop or refine definitions and structures that are missing or have defective definition or structure. In the case of defective definition or structure, recommend improvement in the data development and maintenance processes that created the defective definitions and architectures.[7]

B. Identify Data Definition Quality Measures

1. Identify and, if necessary, define the essential and critical quality characteristics of data definition and information architecture. These are the minimum HUD requirements for

a. Data names, including business term, abbreviated name, database or file name, standard screen name and standard report name;
b. Definition;
c. Valid value set (including value definitions) or reasonable range of values;
d. Business rules for data integrity;
e. Physical specifications of the data.

These quality characteristics must be in place for ensuring effective communication among information producers and consumers or knowledge workers, and data resource management and application development personnel.[8]

C. Assess Data Definition Technical Quality

1. Assess the data definition for conformance to data standards and guidelines. Determine whether the data definition conforms to the minimum established standards.[9] Obtain a comprehensive and concise definition for each data element in the information group. This definition must contain an agreed-upon 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.

2. The data element definition must be known, documented, understood, communicated, and validated by all business areas in the value chain. It should be documented for each record of origin and other metadata for each data element or groups of data elements as described in the illustration in Figure 2.6 below.

D. Assess Information Architecture (IA) and Database Design Quality

1. Assess the information 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, in accordance with three kinds of IA assessments.[10] An *Information Architecture Completeness Assessment* will determine whether the data model has all required entity types and attributes to support the business processes. An *Information Architecture Correctness Assessment* determines whether the data model truly reflects the real world entity types, attributes, and relationships. Finally, an *Information "Chaos" Assessment* determines which instances of data redundancy in Program Area proprietary files, application software package files, or other storage mechanisms are controlled, and which are not controlled.

E. Assess Customer Satisfaction with Data Definition and Information Architecture

1. Measure customer satisfaction with the definition of the information products based on the knowledge workers' assessment. 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.

F. Develop or Improve Data Definitions

1. In cases where the data to be assessed in the subsequent steps lacks or has defective definition and /or information architecture, develop correct definitions and/or information architecture to ensure that subsequent tasks can be executed. Interact with representatives of the business and IT areas across the value and cost chain to arrive at appropriate definitions and architecture (see Section 2-2(E) for a discussion on the identification of the stakeholders).[11]

2. To achieve a new or revised definition, first 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.

3. The template for a data definition as a Business Concept is given in Figure 2.6 below.

Business Concept	Name of the business object to be defined.
Status	[Status] as of x/x/02 [the date of the last revision]
Definition	A succinct, complete business description using common terms or previously defined terms – if new terms are needed, add them to the Business Concept Control Chart and reference in the next section.
Related Business Concepts	• A short-cut or hot-link to a previously defined term. • A description or annotation of an undefined term.
Data Integrity Rules	A bulleted list of data integrity rules (things that "can" or "must" be present for accuracy, completeness).
Unique Identifier	Free form text, bullets, or whatever can be used to provide insights on the appropriate way to uniquely identify each instance of the data entities associated with the business concept.
Life Cycle	Free form text, bullets, or whatever can be used to provide insights on the states and transitions associated with the business concept.
Classification	Free form text, bullets, or whatever can be used to provide insights on the intrinsic classification necessary to understand the business concept.
Domain	• A bulleted list of all applicable domain values. It must be all-inclusive, however if it is not known at the time, a sample can be used while the definition is in process; indicate which approach is used. • Can be a diagram.
Special Usage	Free form text, bullets, or whatever can be used to highlight real-life cases of the use of the business concept.
Examples	Free form text, bullets, or whatever can be used to provide real-life or mock-up illustrations of the business concept
Issues & Concerns	Proposed definition concerns: • A bulleted list of issues or concerns associated with the definition proposed in this document. Existing definition concerns: • A bulleted list of issues or concerns associated with the existing definition, or definitions.
Background Documentation	Free form text and bullets as needed. For specific references to existing document, use the name and date of the document.

Figure 2.6: Illustration of the Business Concept Template

G. Improve Data Development Process

1. 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 is defined with multiple meanings (overloaded data), then the data development and/or data maintenance processes are probably broken. If so, recommend a process improvement initiative to improve the defective process (see Chapter 3).

This improvement must be done prior to the next project requiring new data to be defined and implemented.

2-4 Determine Desired Quality Standards for Prioritized Data Elements

A. Perform in-depth analysis upon the data elements that are within scope. This analysis should identify the information value and cost chain of each data element or group of data elements, describe the element's quality characteristics, and determine the element's quality standard for each quality characteristic.[12]

B. Define Information Value and Cost Chain for Data Element(s)

1. In order to assess data adequately, identify all record(s) of origin for the data. Currently, there are cases where data elements entered in an initial database are updated in a second or even a third system. In cases where redundant data is identified, it must be corrected in every database in which it is stored.

2. The following template (Figure 2.7 below) 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 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

Figure 2.7: Illustration of a Data Element By Record of Origin Worksheet

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

C. Identify Accuracy Verification Sources

1. In order to verify the accuracy of a data element value, it is vital to 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 the potential to contain errors.

D. Determine Applicable Quality Standard for Each Data Element

1. Determine and document the criteria for information quality according to the quality criteria discussed in Section 2-2(F). In setting the desired level of information quality, all criteria should be assessed as to relevance and level of importance. Information that meets the criteria are considered quality; those data elements that do not meet the criteria are considered "defective" and must be corrected or discarded. The example in Figure 2.7 illustrates how to state data element quality criteria.

2. 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 Standards (Compliance Levels)

1. Having defined the quality criteria for each data element, determine what percentage of the data must comply with the specifications. This percentage will be the measuring stick referenced when an organization performs an internal quality assessment. Additionally, the assessment team will use the compliance levels when auditing.

2. 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."

3. Figure 2.8 is an example for documenting the data element compliance targets as well as data exceptions.

Data Element Quality Criteria Worksheet for Record of Origin System: *DQIS (P150)*												
Information Group / 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: both 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: both blank or both not blank.	-	-	-	-	-	Must match the inspection rating on the inspector's log.	-
	Compliance level	95%	-	-	100%	-	-	-	-	-	100% for 1996 and later.	-
	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 HUD 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.	-

Figure 2.8: Illustration of Quality Target Compliance

2-5 Assess Current Level of Information Quality

A. A vital step in information quality improvement is to assess the *current level* of information quality. When selecting the data records for the assessment, acquire a representative, or statistically valid, sample to ensure 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."[13] When properly conducted, a random sample of records provides an accurate picture of the overall information quality of the database.[14] In certain circumstances, purposive samples may be usefully substituted for random samples. For example, if only active cases are of interest, the sample may include all active cases.

B. Extract Random Sample of Data

1. Select a sample size for each point of measure (database, file, transaction flow) based on the following formula:

$$\mathbf{n = ((z \times s) / B)^2}$$

where:

n = the number of records to extract.

z = a constant representing the desired confidence level (e.g., the degree of certainty, expressed as a percentage, of being sure about the estimate of the mean).

s = an estimate of the standard deviation of the data population being measured (e.g., the degree of variation of errors within the data population).

B = the bound or the precision of the measurement (e.g., the variation from the sample mean within which the mean of the total data population is expected to fall given the sample size, confidence level, or standard deviation).

2. Next, generate the random samples. It is critical that data be sampled in a way that ensures the data is unchanged from the point at which the sample is extracted.

3. Provide a sampling report indicating the sample size for each information group, the total universe of data and the portion to be sampled, the sampling criteria and an explanation of the reason why this will produce a statistically valid sample. [15]

C. Measure Information Quality

1. Analyze the information in the samples against its target criteria based on the data definition and information architecture (Section 2-3(F)) as well as the defined quality standards (Sections 2-4(D) and 2-4(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.

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

D. Validate and Refine Data Definitions

1. The data definitions and architectures may be adjusted based on facts discovered during the measurement process. In such cases, the assessment team may execute the information architecture and data definition process described in Section 2-3(F) to arrive at the appropriate revised definitions.

E. Establish Statistical Control

1. For processes that acquire, produce or maintain mission-critical information, it is imperative that they be in a state of statistical control. That is, with respect to the mission-critical information they produce, their results are predictable and the quality of the information is in line with the initial agreed upon levels. Examples of how and when statistical process control of data quality is performed will be available on the EDMG team web site.

2-6 Measure Non-Quality Information Costs

A. The objective of this step is to identify the cost of non-quality information for the information groups or data elements under assessment. Non-quality information costs are assessed in three areas: process failure costs, information scrap and rework costs, and lost or missed opportunity costs.[16]

1. **Process failure costs** – the result of a process, such as distribution of funds, which cannot be accomplished due to missing, inaccurate, incomplete, invalid, or otherwise non-quality information.
2. **Information scrap and rework costs** – incurred when a knowledge worker has to waste time handling or reconciling redundant data, hunting for missing information, verifying data, or working around broken processes.
3. **Lost or missed opportunity costs** – HUD may be missing out on opportunities to greatly improve the lifestyle of communities due to non-quality information, or may be directing funds toward areas of lesser need.

B. Identify Business Performance Measures

1. Information has value to the extent that it enables the enterprise to accomplish its mission or business objectives. In order to determine if a process or information set adds value to the organization, it is important to understand

a. The business vision and mission,
b. The business plans and strategies,
c. The strategic business objectives.

Business performance measures are identified directly from goals set forth in HUD's APP.

C. Calculate Information Costs

1. Identify what percent of information systems and data development and maintenance is value-adding and what percent is cost-adding, performed solely to improve information systems productivity and effectiveness. Determine the relative costs of application and data development in the three categories of information development expense:

   a. **Infrastructure**: These are the costs of developing and reusing databases and applications that create and maintain the information resources.
   b. **Value-adding**: These are the costs associated with the access and retrieval of information for use to add value.
   c. **Cost-adding**: These are the costs of redundancy of both application and data development and maintenance.

D. Calculate Non-Quality Information Direct Costs

1. Quantify the cost of non-quality information to determine the business impact of information quality problems, raise awareness of the importance of information management, and establish a benchmark for measuring information quality initiatives. This process identifies the categories of costs of poor-quality information and calculates the costs of information scrap and rework.

E. Measure Lost Opportunity Costs and Information Value

1. Identify the indirect costs derived from lost or missed opportunities. These can be revenue not realized, missed service to the public, or the costs of underpaying or overpaying for the services rendered. These costs are usually very large and difficult to estimate.

2-7 Interpret and Report Information Quality State

A. Once data has been assessed, the results will be analyzed, interpreted and clearly presented in a format easily understood by knowledge workers, information producers, and process owners. The results will include all assessments of all components (definition, content and presentation). Also, the results will describe findings and recommendations in the quality standards (expected levels of quality), actual quality levels, and information costs, especially non-quality information costs. Each report will include a cover sheet, a summary and detail section, and an assessment procedure report for each information group (see example Figure 2.9 below).[17]

B. The quality assessment summary and detail sections will describe the current level of information quality and then will make recommendations for succeeding 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 Information Quality Assessment Report will describe the following:

1. A description of the approach, such as:
   a. The Source system(s) included.
   b. The Assessment criteria.
   c. The Specific HUD participants.
2. Current level of information quality:
   a. General conclusions about the information quality of assessed elements.
   b. Information quality defects found.

c. The Assessment results (number and types of errors found, level of confidence in the results, any other issues).

3. Recommendations to close the gap between current information quality levels and target information quality standards:
   a. For Data Corrections, indicating appropriate approaches for the errors identified.
   b. For Information Quality Improvements, identifying types of changes that may be made to systems, procedures, or technology to ensure the appropriate level of data quality.
   c. High priority tasks or areas of concern to be addressed first.
   d. Start date.

ASSESSMENT PROCEDURE REPORT

Information Group Name: _____ Date Assessed: _____

Time Period Covered From: _____ To: _____ IQ Analyst: _____

File(s) Sampled: _____ Processes Sampled: _____

Sampling Procedure: □ Representative Random Sample

□ Purposive Selection Sample (Purpose):_____

Sample Size: _____ Sample Percent: _____% **Standard Deviation: _____**

Confidence Level: □ 90% □ 95% □ 99% **Precision: _____**

Assessment type: □ Electronic □ Third-party corroboration

□ Physical to surrogate source □ Survey

□ Physical to real object / event

□ Other: _____

Quality Characteristics 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

□ Usefulness and value-adding

Source: *Improving Data Warehouse and Business Information Quality*, Figure 6.9, 190, adapted for HUD

Figure 2.9: Illustration of an Assessment Procedure Report Template

[THIS PAGE INTENTIONALLY LEFT BLANK]

CHAPTER 3. TIQM® IMPROVEMENT PROCESS

3-1 Overview

A. Information 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 information quality improvement process attacks the *causes* of defective data. Eliminating the causes of data defects and the production of defective data will reduce the need to conduct further costly data correction activities.

B. Maintaining information quality is a continuing effort. Critical to the effectiveness of the procedures is an information quality awareness campaign that motivates information producers and knowledge workers to take daily ownership of information quality. As consumers and producers of quality data, knowledge workers and data providers are the best resource in identifying both quality issues and their solutions.

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

D. The improvement process consists of five major steps (source: *Improving Data Warehouse and Business Information Quality*, p. 289-302):

1. Select Process for Information Quality Improvement.
2. Develop **Plan** for Information Quality Process Improvement.
3. **Do** Implement Information Quality Improvement.
4. **Check** Impact of Information Quality Improvement.
5. **Act** to Standardize Information Quality Improvement.

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

F. The implementation of information quality improvements will include one or more of the following actions:

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

3-2 Select Process for Information Quality Improvement

A. The first step in planning improvements is to identify which process(es) are the best candidates for process improvement. Candidate processes can be identified through the Information Quality Assessment report generated in Section 2-7. The candidate processes

are then prioritized by the best return in information quality for estimated time/cost/effort investment. The return on investment is estimated by reviewing[18]

1. Technical Data Definition Quality Assessment developed in Section 2-3(C),
2. Information Architecture and Database Assessment developed in Section 2-3(D),
3. Information Quality Assessment developed in Section 2-5,
4. Information Cost and Value Analysis developed in Section 2-6,
5. Any data definition Customer Surveys that may have been conducted.

B. Based on the nature of the problem(s) to be solved, a Program Area Quality Improvement Team with representatives from all stages in the value chain is put into place. For an information quality improvement initiative to be effective,

1. Improvement Team representatives must perform the actual work.
2. A non-blaming, non-judgmental environment for process improvement is established. If there are defective data, it is because there are defective processes—not defective people.
3. The process owner or supervisor of the process to be improved must empower the team to make and implement the improvements.
4. 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 possible.
5. A process improvement facilitator must be available and trained in conducting PDCA process improvement.
6. The origin of the data and its downstream uses must be understood.

3-3 Develop **Plan** for Information Quality Process Improvement

A. The foundation for developing information quality procedures is the investigation into the current processes controlling the data and an evaluation of possible root causes. All data control processes must be considered, manual and automated. All sources of the data must be considered, as well as who modifies the data or influences what the data looks like on a form, screen or report.

B. Conduct Root Cause Analysis

1. Once the process is understood, analyze it to get to the "root cause" of a data defect using the Cause and Effect or Fishbone Diagram (see Figure 3.1 below), the "why analysis" technique, or any other method for root cause analysis. Six possible categories of failure cause are included in the Cause and Effect diagram – Human Resources, Material, Machine, Method, Measurement and Environment. For each defect cause identified, it is necessary to answer "why" the error occurred until the root cause is found. All possible scenarios for tracking down the root cause should be explored by considering all six categories in the analysis. A typical root cause analysis might develop as:

2. Scenario - Defect identified is Customer Number not on Order:

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

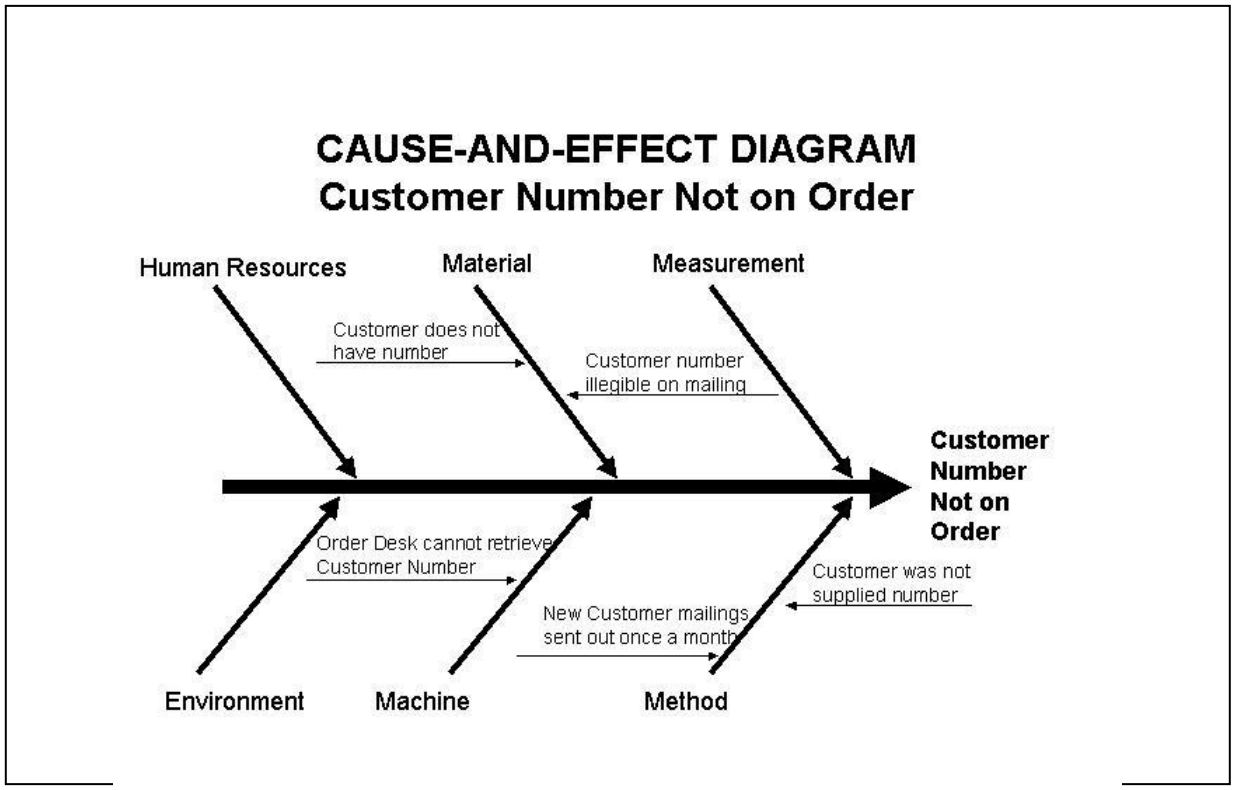

Figure 3.1: Illustration of a Cause-and-Effect Diagram

The diagram in Figure 3.2 presents typical areas of concern needed when applying the Cause-and-Effect diagram to the study of an Information Quality Issue.

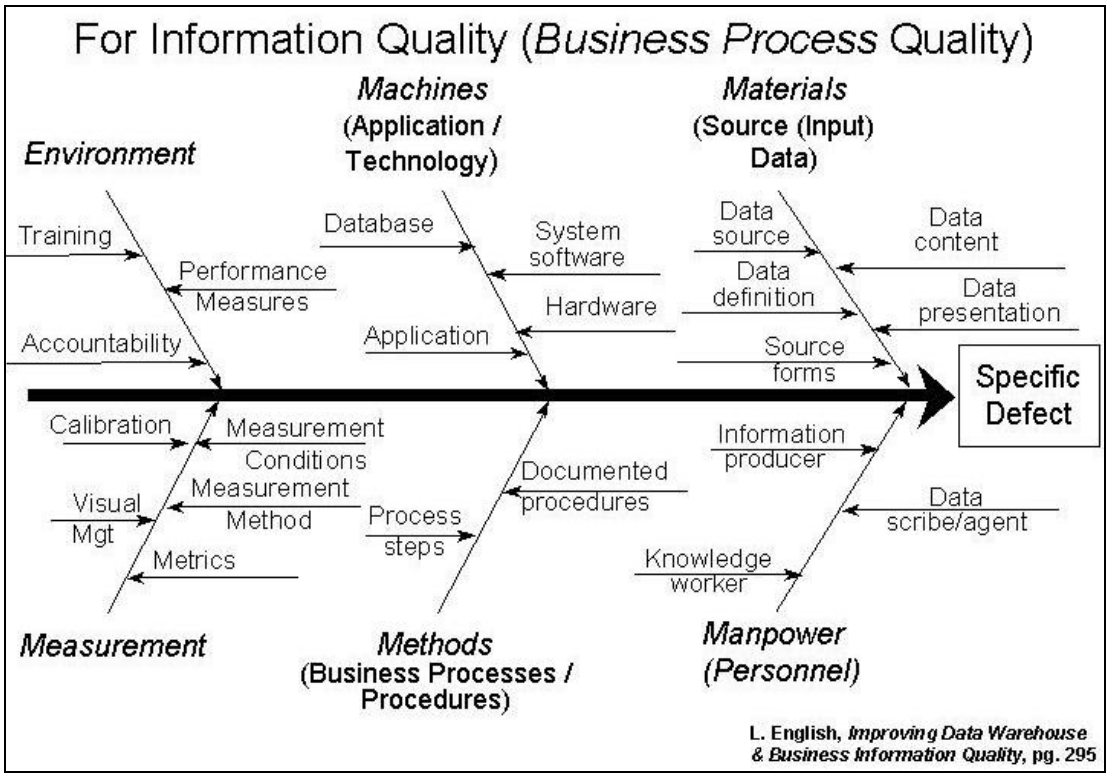

Figure 3.2: Cause-and-Effect Diagram Template for Information Quality

C. Define Process Improvement(s)

1. Process improvement(s) should be defined only *after* root cause(s) have been identified and understood. Otherwise, it may be that only the symptoms or the precipitating cause of the problem are being attacked and not the root causes.

2. Data correction activities can also be leveraged into process improvements.

   a. Any automated correction techniques should become permanent software edits.
   b. Any manual correction procedures should either become permanent software edits or transition to heavily emphasized sections in the information quality awareness and training programs.

3. Just as the categories of failure cause may be Human Resources, Material, Machine, Method, Measurement and Environment, the recommended improvement may involve improvements in any of those categories. Improvements should not be limited to "program fixes." Each of these categories of cause 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 all categories include

   a. Reengineer business processes to include procedures to ensure information quality. For example include supervisor review of critical transactions before entry.
   b. Enforce information stewardship by holding managers and business process owners accountable first, then information producers, for their information quality, completeness and timeliness as well as for other required quality standards.
   c. Allow data element domains to have a value of "unknown" in order to allow information producers to identify an unknown data value rather than entering a guess. The data producer may not know all possible data values, and the "unknown" value may allow for future analysis and correct interpretation.
   d. Identify and designate official record-of-origin, record-of-reference, and authorized record duplication databases.
   e. Adequately train information producers.
   f. Define information quality targets and measures and report information quality regularly.
   g. Implement effective edits that may prevent entry of defective data or flag entry of defective data for later correction.[19]

4. A specific set of information quality standards, procedures and performance measures should be linked to each data control process. Ideally, performance measures should encourage all information producers to create or maintain quality data. The performance should be measured at the time of data capture.

5. Another valuable guideline to help maintain information quality is implementing a single process for data creation and maintenance along with a single application program for each information type, such as stakeholder, address and property. These standard, commonly defined processes and applications should be implemented as early as possible in the information life cycle (or value chain). The databases and data elements must also be standardized to support the information requirements of all information customers.

3-4 **Do** Implement Information Quality Improvement

A. Develop the information quality improvements to implement the recommended solution and implement in a controlled fashion.

1. Document the new procedures, training, software modifications, data model and data base changes as required.
2. Identify a controlled area in which to test the process improvement.
3. Implement the change. If a "people" process is changed, provide orientation and draft procedures. If software changes, deploy the new version to a test area and provide knowledge workers training and/or draft documentation if necessary.[20]

3-5 **Check** Impact of Information Quality Improvement

A. Once the process improvement has been deployed to the test environment, the results of the improvement must be evaluated to verify that it accomplishes the desired quality improvement without creating new problems.

1. Measure and quantify the benefits gained in business performance measures.
2. Quantify economic gains.
3. Record lessons learned.

If the desired results are not achieved without introducing new problems, the implementation of the quality improvement must be adjusted (Section 3-4), or a different solution identified (Section 3-3).

3-6 **Act** to Standardize Information Quality Improvement

A. Once the information quality improvement has been checked, it can replace the old, defective process(es) (source: *Improving Data Warehouse and Business Information Quality*, p. 301-302).

1. Roll out the improvements formally.
   a. Formalize the improved business procedures and documentation
   b. Implement software and database changes into production
2. Implement quality controls as necessary.
3. Communicate to all affected stakeholders.
4. Document:
   a. Lessons learned
   b. Best practices
   c. Cost savings and opportunity gains realized
   d. Process improvement history

[THIS PAGE INTENTIONALLY LEFT BLANK]

CHAPTER 4. TIQM® CORRECTION PROCESS

4-1 Overview

A. This chapter describes the method and techniques that can be used to perform data correction. Unlike information 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 Information Quality Process simultaneously with the Data Correction.[21]

B. Data Correction applies to a variety of efforts such as

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

C. 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 one and the same).

D. 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, all defects identified by the ECTL components must be corrected, captured and reported back to the producing area. This applies whether the correction is one-time (e.g., for historic files) or on-going (files with reference or transaction data provided by the operational systems).

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

4-2 Plan Data Correction

A. Careful planning shortens the time it takes to perform correction and will ensure that resources are available when needed. Establish interim and completion milestones for each task to provide clear indicators of progress and problems.

B. Several planning activities should occur in parallel:

1. Determine appropriate correction approach.
2. Update the correction plan and schedule.
3. Determine automated tool support requirements and schedule.

C. This step will produce the Data Element Correction Plan with the following outline:

1. Identification of correction steps for each data element/data element group.
2. Discussion of the feasibility of data element correction.
   a. Are source documents available?
   b. Is it too costly to correct?
   c. Is a "correct" data element critical to the conduct of business within HUD or with external partners?
3. Description of overall correction approach.
4. Updated Work Breakdown Structure including tasks that:
   a. Identify resources for data element correction.
   b. Identify automated tool support requirements and schedule (Required Deliverable).
5. Deliverables list (Required Deliverable).
6. Updated, detailed correction schedule.

D. Identify and Prioritize Data to be Corrected

1. Using the Information Value and Cost Chain (developed in Section 2-2(D)) and the Information Quality Report (developed in Section 2-5), in conjunction with the Information Value and Cost Analysis (developed in Section 2-6), rank the data by quality, cost to correct, and benefits if corrected.

2. The state of quality, feasibility and cost of correcting must be considered in designing the correction steps for a particular data element or elements.

E. Identify Methods for Data Correction

1. The Information Quality Report 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:

   a. Identification and consolidation of duplicate data.
   b. Correction of erroneous data values.
   c. Supplying of missing data values.
   d. Calculation or recalculation of derived or summary data values.

2. 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. Document and provide in report format a summary of the information defects and the related correction techniques/steps to be applied.

3. Once the appropriate correction steps for each data element or group of similar data elements are documented, fully 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 in the correction process. 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

A. Although the initial assessments detailed in Section 2-5 provide a measure of information 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 if additional data elements can be identified. Analyze and map the data against

the information architecture (Section 2-3 and Section 2-5) to ensure all data elements are identified and fully defined with all associated business rules.

B. Plan and Execute Data Extraction

1. A random sampling of data is extracted from the source database or set of related databases (see Section 2-5(B)). Any method may be used to generate the random sampling, as long as a fully representative sample is produced.

C. Analyze Extracted Data

1. First, the extracted data is parsed down to the atomic level attributes to ensure that all data is examined at the same level. Once parsed, the specific data values are verified against the data definition to identify anomalies. The data is reviewed with subject matter experts to confirm business rules and domain sets, and to define revealed "hidden" data. The data is also reviewed for patterns that may reveal not-yet-documented business rules, which are then also confirmed by the subject matter experts. It is not unusual to find that data, which first appeared anomalous, helps to rediscover forgotten business rules.

D. Document Findings

1. In this step, the definition, domain value sets, and business rules for each data attribute in the database or set of related databases are documented in the Data Definition Worksheet (see Figure 4.1), and the relationship of the data attributes is mapped to the source files and fields using the Data Mapping Worksheet (see Figure 4.2). This information will be used in the transformation process.

Data Definition Worksheet

System: TRACS

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 corresponding 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.

(Source: *Final HUD Data Quality Assessment PAS, LOCCS, HUDCAPS, REMS, TRACS, SAMS, and MTCS Volume 2 dated March 30, 2001. Modified to incorporate revised quality standards*)

Figure 4.1: Illustration of a Data Definition Worksheet

Data Mapping Worksheet		
Data Element	**MTCS Head of Household SSN**	**TRACS 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 equal to "00."
		Value of the last four digits cannot equal "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 not contain a suspicious value of 000000000, 111111111, 222222222, 333333333, 444444444, 555555555, 666666666, 777777777, 888888888, 999999999, 123456789, and 987654321.	Value cannot be equal to "000000000."
		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.

(Source: *Final HUD Data Quality Assessment PAS, LOCCS, HUDCAPS, REMS, TRACS, SAMS, and MTCS Volume 2 dated March 30, 2001. Modified to incorporate revised quality standards*)

Figure 4.2: Illustration of a Data Mapping Worksheet

4-4 Execute Manual and Automated Data Correction

A. In this step the manual and automated corrections are developed, tested, and executed. Data correction tasks include

1. Validating correct data,
2. Correcting erroneous data,
3. Supplying missing data,
4. Consolidating duplicate records (optional depending on IQ effort),
5. Enhancing information with data from external sources (optional depending on IQ effort).

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 re-use of these correction techniques in subsequent projects.

B. Standardize Data for Atomic Level Format and Values

1. The data is examined across databases for consistency as to their definition, domain value, and storage format, use of non-atomic data values, and instances of domain duplicate values (e.g., Sept and Sep).

2. If the data definitions and architectures require refinement based on the actual data in the files and databases, initiate a data definition effort based on the process described in Section 2-3(F). Once the rules for standardization have been reaffirmed, the source data can be mapped against the standardization and the data merge and transformation rules.[22]

C. Correct and Complete Data

1. Correct and complete the data identified in Section 3-2 to the highest quality feasible. This process is particularly significant if the source data is subsequently transformed and enhanced to be incorporated into a data warehouse or data mart. Data anomalies include:

   a. Missing data values.
   b. Invalid data values (out of range or outside of domain value sets).
   c. Data that violates business rules:
      (1) Invalid data pairs (e.g., a Retire Date for an Active employee).
      (2) Superfluous data (e.g., an Employee has two Spouses).
   d. "Suspect data"
      (1) Duplicate data values when unique are expected.
      (2) Overabundance of a value.
      (3) Data that "looks wrong" (e.g., an SSN of 111-11-1111, Start Date of Jan 01, 1900).

2. Occasionally, some data may be "uncorrectable." There are several alternatives to handle this situation.

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

3. 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 less than correct, and incorrect data is often more costly than missing data.

4. Document the method for correcting each data type and the method used for handling uncorrectable data (see Figure 4.3 below). Also, document the cost for correcting each data type to track the expense of information cost and rework. Costs include

   a. Time to develop transformation routines,
   b. Cost of data correction software,
   c. Time spent investigating and correcting data values,
   d. Cost of computer time,
   e. Cost of materials required to validate data.

5. Other, much larger, costs associated with the non-quality information must be identified and quantified such as (source: *The ABCs of Information Quality* seminar; Brentwood, TN; Information Impact International, p. 36-37)

   a. Costs of non-quality information (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 hunting or chasing missing information; 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; recovery costs of unhappy customers.
   b. "Losses" measured in revenue, profit or customer lifetime value, including lost opportunity costs and missed opportunity costs.
   c. Mission failure (Risk) with impact such as the inability to accomplish mission or even to go out of business.

Data Correction Worksheet

System: _____

Data Group (data element list): _____

Correction 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 4.3: Illustration of a Data Correction Worksheet Template

D. Match and Consolidate Data

1. 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 all 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.

2. Match criteria are usually developed 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

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

3. 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.

4. Records with matches are examined 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).

E. Analyze Defect Types

1. The errors identified in the previous steps are analyzed 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.

2. These results are compiled in the Data Element Correction Summary Report with the following outline:

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

F. Transform and Enhance Data

1. Once the data has been corrected, prepare for loading back to the source database or into the target database. In the cases where data transformation is required, the transformation process addresses any data conversions necessary as identified in

Section 4-4(B). The enhancement process augments internal data with data from an external data source.

2. 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 the data transformations expected include the following:

   a. **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.
   b. **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 month as numbers 1-12 may require a conversion to the three-character code.
   c. **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, where 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.
   d. **Vertical filter:** A field used for multiple purposes is split into discrete fields for each purpose.
   e. **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.
   f. **Matching and consolidation**: Records identified in Section 4-4(D) above and verified as true duplicates are consolidated.
   g. **Data evaluation and selection:** As records are combined from multiple data sources to a data warehouse or other database, select the most authoritative data. 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 or behavioral and census data from an external source to support an identified business need. For example, income information may be obtained from an external source and appended to client records to help determine their Section 8 benefits.

G. Calculate Derived and Summary Data

1. If data is summarized or derived, calculate this data. This usually applies to a data warehouse or data mart ECTL. Data is summarized or combined to optimize performance for frequent queries against the database. This can be accomplished through the following steps:

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

4-5 Determine Adequacy of Correction

A. Before the project can be brought to a close, the success of the correction process must be evaluated. At a minimum, the following checks should be performed (adapted from *Improving Data Warehouse and Business Information Quality*, p. 275-278):

1. Determine each data element's post-correction quality compliance level. Check a sample to verify:
   a. Data values fall with the domain value set or range, if any.
   b. "Missing" data values are now present.
   c. Data values follow business rules.
   d. Data is loading according to specified data mapping (as developed in Section 4-4(B)).
2. Verify effects of transformation and enhancement. Again, check output results to verify:
   a. Transforms performed as expected.
   b. Records are enhanced with the correct data as expected.
3. Verify all records are loaded as expected.
   a. All jobs ran to completion.
   b. Correct number of records were processed.
   c. None of the records were inadvertently processed twice.
   d. Correct number of duplicate records consolidated.
4. Document the impact of the correction techniques, percent of errors or omissions
   a. Corrected accurately using automated means.
   b. Corrected through human efforts or means.
   c. Corrected to an inaccurate value (valid, but not accurate).
   d. Not corrected because it was impossible or cost prohibitive to get the correct value.
5. Document which correction techniques worked and which did not work.
6. Analyze the information defects and recommend appropriate improvements.
7. Update the Data Element Quality Criteria Worksheet (Figure 2.8).
8. Document adjustments to the correction schedule.

B. Produce the Data Element Correction Adequacy Report with the following outline:

1. An assessment of correction techniques, especially which techniques should be re-used.
2. Determination of data element post-correction compliance levels.
3. Summary of improvement in information quality.
4. Analysis of IQ weaknesses and recommendation of corresponding improvements.
5. Assessment of correction plan, Work Breakdown Structure, schedule, required human resources, and roles.
6. Identification of next steps.

[THIS PAGE INTENTIONALLY LEFT BLANK]

CHAPTER 5. TIQM® CERTIFICATION PROCESS

5-1 Overview

A. This chapter describes the method and techniques that will be used by the EDMG to perform the final task of independent verification or certification of mission-critical information. This task of independent verification, or certification, takes place after the processes that produce or maintain selected data elements and information groups are improved and the existing data has been corrected. Based on the established priorities and schedules, the EDMG will verify that the level of data quality achieved is aligned with the expectations of all business areas that consume the information. This certification will be in two areas:

1. 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.
2. Second, to assess whether the data contained in files, databases, data warehouses, data marts, reports, and screens are also in compliance. This assessment will help evaluate the adequacy of data correction efforts.

B. Based on its observations and findings, the EDMG will recommend improvements to the procedures used to implement information quality improvements (defect prevention) as well as improvements in data correction procedures. In addition, if the certification process finds shortfalls in information quality, the responsible Program Areas will need to submit a new schedule and perform additional information improvement and/or correction.

5-2 Certify Information Quality Process Improvements

A. This activity is similar to the "Check Impact of Information Quality Improvement" activity, described in Section 3-5. Before a meaningful certification of an information process improvement can be performed, the process under improvement must be certified as being in statistical control. That is, the process must be producing a consistent (predictable) and acceptable level of quality of information (the data is consistently meeting all knowledge workers and end customer needs). Once the process is in statistical control, it is possible to determine that the changes indeed produced the expected improvements.

B. Verify the effectiveness of the Information Quality Improvement process by assessing the results of the information quality improvement. Critical points to be assessed include

1. Was the information quality improvement planned appropriately? Is there something that can be done to improve the process? The plans (the "P" in the PDCA cycle) and the actual execution logs will be used to determine if the process needs to be revised for improvement.
2. Was the information 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 (the "D" in the PDCA cycle).
3. Were the information quality improvement results checked for impact across the information value chain? This is the process of determining the effectiveness of the "check" (the "C" in the PDCA cycle).

4. Were the actions to standardize the information quality improvement across the target environment effective? Were the expected results achieved? The actual rollout or "Act" (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 EDMG 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

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

B. Define the Scope of the Certification

1. 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 the Program Area has identified as ready for certification and for the assessed information 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 indicates 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.

C. Identify the Data Element Definitions

1. If the Program Area has applied the TIQM® approach described in this Handbook, then the comprehensive definition will already be specified for each data element. Refer to Section 2-3 for a detailed discussion on this task. However, if the data definition is not in place, it will be defined using the approach described in Section 2-3(F).

D. Define Certification Approach

1. Based on the prior assessment for each information group, determine one or more techniques for assessing their actual level of quality. Refer to Section 2-5(C) for details on this selection.

E. Define Sample Size and Resources

1. Based on the prior assessment, for each information group and for each assessment point, determine the sample size using the same approach as the prior assessment (see Section 2-5(B)). Identify the participants in the assessment process and the estimated number of hours and calendar days required. Identify any special requirements, such as access to documents, acquisition of tool(s) not already in HUD's inventory, travel.

F. Develop Certification Criteria

1. The terms of the certification will be the same as those agreed upon as part of the original assessment, unless otherwise agreed to by the OCIO and the Program

Area, based on lessons learned during the data correction process or special conditions identified by either party.

G. Conduct Certification

1. Perform the tasks in the certification plan to determine the level of compliance of the data elements within the scope of the certification.

H. Interpret and Report Information Quality Certification

1. Once the data has been certified, report the results as stated in Section 2-7, replacing the term “assessment” with the term “certification.”

[THIS PAGE INTENTIONALLY LEFT BLANK]

APPENDIX A. TIQM® PLANNING

The HUD TIQM® method defines four major processes. However, information quality assessment or certification projects most frequently will consist of one TIQM® process, whereas improvement and correction projects will be more typically a combination of these two processes. In some cases, projects may add steps or tasks within processes to meet the particular needs of the Program Area.

The decision as to 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 necessarily be updated at key points throughout the project and the level of detail of the planning 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 TIQM® tasks may be iterative based upon the requirements of the individual Program Areas and the state of the information quality. These samples were developed to help identify the high level tasks required to plan and execute a TIQM® 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 TIQM®

A project plan typically includes the components listed below. The TIQM® Project Plan for an Improvement or a Correction project is a required document. However, only the project Schedule is a required deliverable.[23]

- 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 all items to be delivered plus 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 TIQM® 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 TIQM® ASSESSMENT WORK BREAKDOWN STRUCTURE

1.0 Plan TIQM® Assessment Project

2.0 Select Data Element Candidates

- 2.1 Determine Scope Based on Business Needs
- 2.2 Identify Information Group to be Assessed
- 2.3 Identify Information Value and Cost Chain
- 2.4 Identify Information Stakeholders
- 2.5 Identify Information 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 Information Architecture Quality

- 3.1 Identify Data Definition Quality Measures
- 3.2 Assess Data Definition Technical Quality
- 3.3 Assess Information Architecture and Database Design Quality
- 3.4 Assess Customer Satisfaction with Data Definition and Information 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 Information Value and Cost Chain 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)
- 4.5 Determine Presentation Quality Measures
- 4.6 Establish Statistical Controls

5.0 Assess Current Level of Information Quality

- 5.1 Extract Random Sample of Data
- 5.2 Measure Information Quality

6.0 Measure Non-Quality Information Costs

- 6.1 Identify Business Performance Measures
- 6.2 Calculate Information Costs
- 6.3 Calculate Non-Quality Information Costs
- 6.4 Measure Lost Opportunity Costs and Information Value

7.0 Interpret and Report Information Quality

A.3 SAMPLE TIQM® IMPROVEMENT WORK BREAKDOWN STRUCTURE

1.0 Plan TIQM® Improvement Project

2.0 Implement Information Quality Improvement

- 2.1 Prepare for Information Quality Improvement
  - *2.1.1 Select Process for Information Quality Improvement*
  - *2.1.2 Develop Information Quality Awareness Plan*
  - *2.1.3 Identify PAQIT Members*
  - *2.1.4 Assess Skills Levels of Team Members*
  - *2.1.5 Conduct Information Quality Improvement Training as Needed*
- 2.2 Plan for Information Quality Process Improvement
  - *2.2.1 Determine Information Stewardship Roles Across the Value and Cost Chain*
  - *2.2.2 Conduct Root Cause Analysis*
  - *2.2.3 Define Process Improvement(s)*
  - *2.2.4 Develop Information Quality Improvement Measures*
  - *2.2.5 Determine Improvement Strategy, Plan, Milestones, And Schedule*
- 2.3 Do Implement Information Quality Improvement
  - *2.3.1 Document Improved Procedures and Training*
  - *2.3.2 Document Improvement Changes for Software, Data Models and Databases*
  - *2.3.3 Develop Procedural Modifications and Train Knowledge Workers*

2.3.4 Develop Application Software Edits
2.3.5 Identify a Controlled Area for Deployment
2.3.6 Implement in a Controlled Environment
2.4 Check Impact of Information Quality Improvement
2.4.1 Measure and Quantify Performance Benefits
2.4.2 Quantify Economic Gains
2.4.3 Record Lessons Learned
2.5 Act to Standardize Information Quality Improvement
2.5.1 Implement Necessary Quality Controls
2.5.2 Deploy Improvements
2.5.3 Record Lessons Learned and Best Practices
2.5.4 Record Costs Savings, Opportunities Realized
2.5.5 Record Process Improvement History

A.4 SAMPLE TIQM® CORRECTION WORK BREAKDOWN STRUCTURE

1.0 Plan TIQM® 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 Determine Appropriate Data Correction Techniques and Methods
2.1.4 Determine Automated Tool Support Requirement
2.1.5 Procure and Install Automated Tools as Required
2.2 Extract and Analyze Source Data
2.2.1 Identify Sampling Technique and Extract Sample
2.2.2 Analyze Data
2.2.3 Document and Confirm Findings
2.3 Execute Manual and Automated Data Correction
2.3.1 Develop and Test Automated Data Correction Techniques
2.3.2 Develop and Test Manual Data Correction Procedures
2.3.3 Execute Data Correction
2.3.4 Summarize Data Correction Activities
2.3.5 Calculate Derived and Summary Data
2.3.6 Verify Corrections Properly Applied (Procedure)
2.3.7 Summarize Data Correction Activities
2.4 Determine Adequacy of Correction
2.4.1 Assess Correction Techniques and Re-Usability
2.4.2 Identify the Data Elements/Groups Corrected (Content)
2.4.3 Re-Assess Data Element Quality
2.4.4 Assess Plan, Schedule, and Resource Roles
2.4.5 Assess Cost Effectiveness
2.4.6 Review and, as Necessary, Redefine Quality Criteria Specifications
2.4.7 Identify Best Correction Practices
2.4.8 Recommend Improvements to Correction Tasks

A.5 SAMPLE TIQM® CERTIFICATION WORK BREAKDOWN STRUCTURE

1.0 Plan TIQM® Project

1.1 Identify Information Quality Improvement for Certification
1.2 Revise Terms and Conditions for Certification
1.3 Develop Certification Plan

2.0 Certify Approach

2.1 Select Information Group Candidates
2.1.1 Validate Information Value and Cost Chain
2.1.2 Identify Information Stakeholders
2.1.3 Identify Information Quality Objectives and Measures

2.1.4 Determine Files and Processes to Certify
2.1.5 Prioritize Data Elements Supporting Business Needs
2.2 Validate Data Definition and Information Architecture Quality
2.2.1 Validate Data Definition Quality Measures
2.2.2 Assess Data Definition Technical Quality
2.2.3 Assess Information Architecture & Database Design Quality
2.2.4 Assess Customer Satisfaction with Data Definition and Information Architecture
2.2.5 Prioritize Data Elements Supporting Business Need
2.3 Measure Non-Quality Information Costs
2.3.1 Identify Business Performance Measures
2.3.2 Calculate Information Costs
2.3.3 Calculate Non-Quality Information Costs
2.3.4 Measure Lost Opportunity Costs and Information Value

3.0 Certify Information Quality Process Improvements

3.1 Define Information Quality Improvement Process Certification Technique
3.2 Identify Stakeholders (Interviewees)
3.3 Develop Information Quality Improvement Process Certification Criteria
3.4 Conduct and Manage Information Quality Improvement process Certification
3.5 Interpret and Report Information Quality Improvement Findings

4.0 Certify Data Corrections

4.1 Identify Data Element Definitions
4.2 Define Data Certification Technique
4.3 Define Data Certification Technique
4.4 Define Sample Size and Resources
4.5 Develop Data Certification Criteria
4.6 Conduct and Manage Data Certification
4.7 Interpret and Report Findings

APPENDIX B. INFORMATION QUALITY SOFTWARE TOOLS

Information quality tools provide automation and management support for solving information quality problems (source: *Improving Data Warehouse and Business Information Quality*, Chapter 10). Effective use of information quality tools requires

- Understanding the problem you are solving,
- 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 your requirements,
- Using the tools properly.

Sections B.1-B.5 below discuss five categories of information tools for information quality improvement and data correction that may be applied within individual Program Areas at HUD to support the four-stage process for information quality improvement discussed in this Handbook. It is recommended that each Program Area choose its own tools to support specific program business needs. It is always recommended that a Program Area select only one tool to accomplish a single category of information quality improvement.

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

B.1 INFORMATION 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 type 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 affecting the data. This analysis may identify quantitative (formula-based) or qualitative (relationship-based) conditions affecting the data and its successful migration and transformation. The analysis may also uncover exceptions or errors in the conditions.

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 fraud, intentional or unintentional.

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 analysis 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 in-house data knowledge and analysis to adequately use them. Data correction tools may be used to standardize data, identify data duplication, and transform data into a correct set of values. These tools are invaluable in automating the most tedious facets of data correction.

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

- Extracting data.
- Standardizing data.
- Matching and consolidating duplicate data.
- Reengineering data into architected data structures.
- Filling in missing data, based upon algorithms or data matching.
- Applying updated data, such as address corrections from change of address notifications.
- Transforming data values from one domain set to another.
- Transforming data from one data type to another.
- Calculating derived and summary data.
- Enhancing data, by matching and integrating data from external sources.
- Loading data into a target data architecture.

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 problems before they occur.

Defect prevention tools provide the same kind of functions as data correction tools. The difference is that they provide for discovery of the errors and correction of them 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 data definition and information architecture development. The tools perform one or more of the following functions:

- Ensure conformance to data naming standards.
- Validate data name abbreviations.
- Ensure all required components of data definition are provided.
- Maintain metadata for control of 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 information product. These tools cannot determine if data required for knowledge workers is missing, defined correctly, or even required in the first place. Information resource data (metadata) quality tools may audit or ensure that data names and abbreviations conform to standards, but they cannot assess whether the data standards are “good” standards that produce data names that are understandable to knowledge workers.

B.6 EVALUATING INFORMATION QUALITY TOOLS

Tool selection is second only to the business problem at hand in architecting a business solutions environment. Evaluate any software tool from the standpoint of how well it solves business problems and supports accomplishing the enterprise business objectives. Avoid "vendor pressure" to buy tools *before* you develop requirements.

Once you understand the business problems you are solving, determine what category of information quality function automation is required. For example, the fact that you are developing a data warehouse does not automatically mean that your 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 information producers do not know who uses the information they create. Therefore, a data defect prevention tool is required to solve the real business problem.

[THIS PAGE INTENTIONALLY LEFT BLANK]

APPENDIX C. TOTAL INFORMATION QUALITY MANAGEMENT CONCEPTS

The following sections present background necessary to understand the evolution of thought in quality management. Section C.1 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. Section C.2 gives a proper understanding of quality standards based on customer needs and expectations and not on artificial goals.

C.1 EVOLUTION OF QUALITY – FROM INTRINSIC TO CUSTOMER CENTRIC

The United States manufacturing industry operated in a steady state from the end of World War II until the late 1970's, when it suffered a revolution caused by a redefinition of quality. The new paradigm of quality owed its creation to the Japanese manufacturing industry's application of Dr. W. E. Deming's principles of quality. Before this revolution, quality was thought to be "product intrinsic" and therefore achievable by after-the-fact inspection (the "quality control" school of thought). If the product was defective, it was either sent back for correction (re-worked) or disposed of (scrapped). However, this approach directly increased costs in three ways: first, the added cost of inspection; second, the cost of re-work; third, the cost of scrap. In those cases where inspection was based on samples (not 100% inspections), there were also the costs of delivering a defective product to a customer (including dissatisfaction and handling of returns). Dr. Deming questioned the quality control approach and affirmed that quality can best be achieved by designing it into a product and not by inspecting defects out of a finished product. He indicated that inspection should be used at a minimum and only to determine if there is unacceptable variability, and advocated a focus on improving the process in order to improve the product. Also, he centered his definition of quality on the customer, not the product. He indicated that quality is best measured by how well the product meets the needs of the customer.

Dr. Deming's approach, used since the early 1960's,[24] was also based on the "PDCA" approach (continuous process improvement) developed by W. Shewhart,[25] and the Total Quality Management approach developed by P. B. Crosby.[26] 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.[27]

C.2 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."[28] "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.[29]

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."[30]

The same is true for information quality. Larry English's analysis concludes that the costs of non-quality information can be as much as 10 to 25 percent of operating budgets and can be even higher in information intensive organizations.[31] In the absence of a set information quality standard, the standard is as simple: "If information is 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 information you create if that data is needed for other processes."[32]

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 information, both short and long term, and are signed and monitored by both the providers and customers of the information. 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 its 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 D. GLOSSARY[33]

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. **LPE**

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. **LPE**

Accessibility: the degree to which the knowledge worker or end customer is able to access or get the information they need. **LPE**

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

Accuracy to surrogate source: A measure of the degree to which data agrees 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. **LPE**

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

Automated information quality assessment: Information 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 information quality. However, they may not automate assessment of some very complex business rules and they *cannot* evaluate accuracy. **LPE**

Business concept: A person, place, thing, event or idea that is relevant to the business and for which the enterprise collects, stores, and applies information. 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 information steward: The person or group that manages the development, approval, and use of data within a specified functional area, ensuring that it can be used to satisfy business data requirements throughout the organization.

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

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

(HUD's) Common Data Element Correction Method: A method for correcting data, developed for HUD, based upon accepted industry standards and incorporating project management and total quality management principles. Now replaced with HUD's Total Information Quality Management methodology. See *Total Information Quality Management*.

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

Completeness: A characteristic of information quality measuring the degree to which all required data is known. (1) *Fact* completeness is a measure of data definition quality expressed as a percentage of the attributes about an entity type that need to be known to assure 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 columns or fields of a table or file that should have values in them, in fact do so. For example, "95 percent of the columns for the customer table have a value in them." 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 in which HUD needs to know about, have a record (row) in the household table." **LPE**

Concurrency: A characteristic of information quality 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 information float time from when data is available in one data source and when it becomes 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 *Information float*). [LPE]

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

Contextual clarity: the degree to which information presentation enables the knowledge worker or end customer to understand the meaning of the data and avoid misinterpretation. [LPE]

Controllable mission-critical information: Controllable means that HUD has control over information or data content because it is collected following HUD standards (e.g., housing authority filings) or produced within HUD. Non-controllable mission-critical information is information acquired by HUD from sources that cannot be controlled by HUD such as survey information from an external source (e.g., homeownership rates from the Bureau of Census) or information resulting from HUD actions such as surveys from small samples with large margin of error, non-respondents impacting the representativeness of the sample or inaccurate responses from respondents.

Data: The representation of facts. Data can represent facts in many mediums or forms including digital, textual, numerical, or graphical. The raw material from which information can is produced when it is put in context that gives it meaning. [LPE]

Data correction: See *Information Product Improvement*.

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

Data dictionary: A repository of information (metadata) defining and describing the data resource. A repository containing 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 information or metadata. An *integrated* data dictionary is one that is capable of controlling the data and process environments. A *passive* data dictionary is one that is capable of storing metadata or data about the data resource, but is not capable of interacting with or controlling the computerized environment external to the data dictionary. See also *Repository*. [LPE]

Data element: The smallest unit of named data that has meaning to a knowledge worker. A data element is the implementation of an attribute. Synonymous with data item and *field*. [LPE]

Data improvement: See *Information 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). [LPE]

Data quality: See *Information quality*.

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

Data standardization: See *Data Definition*.

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

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

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 information 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 captured.

Derivation integrity: the correctness with which derived data is calculated from its base data.

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

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

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 information; and responses to requests for agency records under the Freedom of Information Act, the Privacy Act, the Federal Advisory Committee Act or other similar law. This definition also does not include distribution limited to correspondence with individuals or persons, press releases, archival records, public filings, subpoenas or adjudicative processes. **OMB**

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.**LPE**

External partner: These are individuals and organizations that provide to and/or receive from HUD services and/or information regarding housing and/or urban development. They include state and local governments, other federal agencies, housing authorities, and public service organizations.

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

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.

Influential information: is scientific, financial, or statistical information that the U. S. Government Agency can reasonably determine that dissemination of the information will have or does have a clear and substantial impact on important public policies or important private sector decisions. **OMB**

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) which 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 information –*specifically* a numerical quantity that measures the uncertainty in the outcome of an experiment to be performed. **MWD**

Information (2): (In the context of business and government use; disseminated or not; this is the definition used in this Handbook) 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. **LPE**

Information (3): (In the context of information 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 information that an agency ***disseminates*** from a web page, but does not include the provision of hyperlinks to information that others disseminate. This definition does not include opinions, where the agency's presentation makes it clear that what is being offered is someone's opinion rather than fact or the agency's views. **OMB**

Information architecture: A "blueprint" of an enterprise expressed in terms of a business process model, showing what the enterprise does; an enterprise information model, showing what information resources are required; and a business information model, showing the relationships of the processes and information. **LPE**

Information dissemination: see *Dissemination of information*.

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

Information group: A relatively small and cohesive collection of information, 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 and several business roles that use the information. **LPE**

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

meet all information stakeholders' needs. See also *Data intermediary*. [LPE]

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 *information scrap and rework*. See also *Data reengineering*. Information product improvement is *reactive* information quality. [LPE]

Information quality: The degree to which information consistently meets the requirements and expectations of the knowledge workers in performing their jobs. [LPE]

Information quality (assessed level): The measurement of actual quality of a set of information against its required quality characteristics. [LPE]

Information quality (desired level): The level of data quality required to support the business needs of all information consumers. [LPE]

Information quality assessment: The random sampling of a collection of data and testing it against its valid data values to determine its accuracy and reliability. Also called *data quality assessment* or *data audit*. [LPE]

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

Information steward: There are seven business roles in information stewardship and nine information systems roles in information stewardship. See *Business information steward*. [LPE]

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

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

Knowledge worker: The role of individuals in which they use information 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 an *information consumer* or *customer*. Accountable for work results created as a result of the use of information and for adhering to any policies governing the security, privacy, and confidentiality of the information used. The term knowledge worker was created by and has been used consistently by Peter Drucker since as early as 1973 to describe in general all "workers" in the Information Age organization. [LPE]

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

Mission-critical information: Is information considered fundamental for HUD to conduct business, or information frequently used by the Department, particularly financial information, key to the Department's integrity and accountability, and information used to support Annual Performance Plan reports. Program Areas, the Office of the Chief Financial Officer (OCFO), the Office of the Chief Information Officer (OCIO), the Deputy Secretary or the Secretary will categorize information as mission-critical. Mission-critical information will be managed using the TIQM® approach to enable HUD to achieve expected levels of information quality necessary to serve its constituents properly (also, see controllable mission-critical information).

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

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 a scientific, financial, or statistical context, the supporting data 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, data should have full, accurate, transparent documentation, and error sources affecting data quality should be identified and disclosed to information consumers. [OMB]

Physical Information Quality Assessment: Physical assessments compare data values to the real-world objects and events that the data represents in order to confirm that the values are accurate. This type of testing is more time and labor intensive than automated testing, but 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. **LPE**

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

Primary key: The attribute(s) 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. **LPE**

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

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

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

Ranges, reasonability tests: General tests applied to information to determine if the value is 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 Patient Temperature might be 80-110 degrees, while the range for Room Temperature might be –20 to 120 degrees.

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

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

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. **LPE**

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

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

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. **LPE**

Scalability: The ability to scale to support larger or smaller volumes of data and more or less knowledge workers. 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 information from which the data was originally entered. Also, an electronic copy of the data generated outside the organization that is known to be accurate. **LPE**

Timeliness: A characteristic of information quality measuring the degree to which data is available when knowledge workers or processes require it. **LPE**

(HUD's) Total Information Quality Management (TIQM®): Techniques, methods and management principles that provide for continuous improvement to the information processes of an enterprise. A management approach used by HUD, based upon accepted industry standards and incorporating project management and total quality management principles. It replaced HUD's Common Data Element Correction Method.

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

User: A term used by many to refer to the role of people in information technology, computer systems, or data. The term is inappropriate to describe the role of information producers and knowledge workers who perform the value work of the enterprise, and for whom information technology should enable them to transform their work and who depend on information to perform their work. With respect to information technology, applications, and data, the role of business personnel is that of information producers and knowledge workers. The term knowledge worker was created by and has been used consistently by Peter Drucker since as early as 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 together solve the information and work problems of the enterprise. **LPE**

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

Validation (of Performance Data): is the assessment of whether the data are appropriate for the performance measure. Also, validation indicates the appropriateness of … performance measures in relation to … goals and objectives. **GAO**

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

Verification (of Performance Data): The assessment of [performance] data completeness, accuracy, and consistency, timeliness, and related quality control practices. **GAO**

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

APPENDIX E. FOOTNOTES

[1] Particularly 2001-FO-0004.
[2] Larry P. English, *Improving Data Warehouse and Business Information Quality (IDW&BIQ)*. (New York: John Wiley & Sons, 1999), 17.
[3] Ibid., 27-30.
[4] Ibid., and Larry P. English, *Information Stewardship: Implementing Accountability for the Information Resource* seminar. (Brentwood, TN, 1993-2002), 21.0.
[5] English, *IDW&BIQ*, 123-124.
[6] Ibid., 123.
[7] Ibid., 119-135.
[8] Cf. ibid., 83-118 for additional discussion of the critical quality characteristics; 119-123 for additional explanation of the procedures and techniques for this task.
[9] For further explanation of procedures and techniques that can be applied in this task, cf. ibid., 126-129.
[10] For an explanation of the procedures and techniques for this task, cf. ibid., 130-133.
[11] For an explanation of the characteristics, cf. ibid., 83-118; for an explanation of the procedures for this task, 119-123.
[12] Ibid., 155-196.
[13] Ibid., 171.
[14] Ibid., 167-188.
[15] To learn more on how to determine an appropriate sample size, confidence level and confidence interval, cf. ibid., 167-177.
[16] Ibid., 213-234.
[17] Ibid., 188-196.
[18] Ibid., 290.
[19] For more examples of Best Practices techniques for process improvements, cf. ibid., 302-309.
[20] For more information, cf. ibid., 298-299.
[21] Ibid., 237-283.
[22] For more information, cf. ibid., 252-257.
[23] Ibid., 188-302; adapted for HUD.
[24] Cf. W. E. Deming, *Out of the Crisis*: Cambridge: MIT Center for Advanced Engineering Study, 1986.
[25] Cf. W. Shewhart, *Statistical Method from the Viewpoint of Quality Control*: New York: Dover Publications, 1986.
[26] Cf. P. B. Crosby, *Quality is Free: The Art of Making Quality Certain*: New York: Penguin Group, 1979.
[27] Cf. M. Imai, "Kaizen: The Key to Japan's Competitive Success:" New York, Random House, 1989; and "Gemba Kaizen: Low Cost Approach to Management:" New York: McGraw-Hill, 1997.
[28] Crosby, *op. cit*., p. 58.
[29] Ibid., 149.
[30] Ibid., 146-147.
[31] English, *IDW&BIQ*, 12.
[32] Larry English, *Information Quality Management: What Managers Must Know and Do* seminar. Brentwood, TN: Information Impact International, 1998-2002, p. 1.2
[33] Glossary definitions are taken from four sources, indicated by a letter superscript following each term: **LPE** = Larry P. English, *Improving Data Warehouse and Business Information Quality*. (New York: John Wiley & Sons, 1999), Appendix A; **MWD** = Merriam-Webster Dictionary; **OMB** = OMB Section 515 or 66-FR-49718; **GAO** = Performance Plans: Selected Approaches for Verification and Validation of Agency Performance Information, Page 12, GAO/GGD-99-139.