

## 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.<sup>21</sup>
- 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.

<b>Data Definition Worksheet</b>	
<b>System:</b> <u>TRACS</u>	
<b>Data Element Storage Details:</b>	
<b>Table:</b> <u>Voucher</u>	<b>Column:</b> <u>Contract Number</u>
<b>Storage Format:</b> <u>Text</u>	<b>Length:</b> <u>10</u>
<b>Definition:</b> <u>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.</u>	
<b>Domain Values:</b> <u>N/A</u>	
<b>Business Rules:</b>	
1. <u>Value contains letters and numbers only.</u>	
2. <u>If value begins with a letter, then value must be a two letter combination corresponding to a valid state code.</u>	
3. <u>If the subsidy type is 1, 7, 8, or 9, then a value must be present.</u>	
4. <u>If the subsidy type is 2, 3, 4, or 5, then a value must NOT be present.</u>	

(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.<sup>22</sup>

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.

<b>Data Correction Worksheet</b>	
<b>System:</b> _____	
<b>Data Group (data element list):</b> _____	
<b>Correction Method Used:</b> _____	
<b>Expenses:</b> _____	
<b>Time Investigating Data Defects:</b> _____	Man Days/Months/Years @ \$ _____ avg. cost
<b>GOTS/COTS Data Correction Software Cost:</b> _____	
<b>Time Spent Correcting Data Values:</b> _____	Man Days/Months/ Years @ \$ _____ avg. cost
<b>Time to Develop Transformation Routines:</b> _____	Man Days/Months/ Years @ \$ _____ avg. cost
<b>Cost of Computer Time:</b> _____	
<b>Cost of Materials to Validate Data:</b> _____	
<b>Total Costs:</b> _____	_____

**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]