UNIVERSITY OF OREGON
APPLIED INFORMATION MANAGEMENT

Presented to the Interdisciplinary
Studies Program:
Applied Information Management
and the Graduate School of the
University of Oregon
in partial fulfillment of the
requirement for the degree of
Master of Science

Improving the Data Warehouse with Selected Data Quality Techniques: Metadata Management, Data Cleansing and Information Stewardship

CAPSTONE REPORT

Brian Evans
IT Business Systems Analyst
Mentor Graphics Corporation

University of Oregon
Applied Information
Management
Program

December 2005

722 SW Second Avenue
Suite 230
Portland, OR 97204
(800) 824-2714

Approved by

____

Dr. Linda F. Ettinger

Academic Director, AIM Program

ABSTRACT

Improving the Data Warehouse With Selected Data Quality Techniques: Metadata Management, Data Cleansing and Information Stewardship

The corporate data warehouse provides strategic information to support decision-making (Kimball, et al., 1998). High quality data may be the most important factor for data warehouse success (Loshin, 2003). This study examines three data management techniques that improve data quality: metadata management, data cleansing, and information stewardship. Content analysis of 14 references, published between 1992 and 2004, results in lists of themes, synonyms, and definitions for each technique, designed for data warehouse analysts and developers.

Table of Contents

Chapter I. Purpose of the Study 1

Brief Purpose 1

Full Purpose 3
Significance of the Study 8
Limitations to the Research 9

Definitions 12

Problem Area 16

Chapter II. Review of References 19

Key literature supporting the research method 19

Key literature concerning enterprise data quality 20

Key literature concerning data warehousing 25

Key literature concerning data quality for data warehouses 27

Chapter III. Method 31

Research question 32

Data collection – Choosing a sample of literature 33

Data analysis - Coding text into manageable content categories 34

Data presentation – Analyzing results 36

Chapter IV. Analysis of Data 39

Code the texts 39
Key literature concerning enterprise data quality 39
Key literature concerning data warehousing 40
Key literature concerning data quality for data warehouses 40

Present the results 45

Chapter V. Conclusions 47

Metadata Management 47
Clarification of terminology 47
Profile of metadata management to assist in the assessment of tools 48
Assistance in the design of metadata management for a data warehouse ... 48

Data Cleansing 49
Clarification of terminology 49
Profile of data cleansing to assist in the assessment of tools 50
Assistance in the design of data cleansing for a data warehouse 50

Information Stewardship 51
Clarification of terminology 51
Profile of information stewardship to assist in the assessment of tools 52
Assistance in the design of information stewardship for a data warehouse 52

Appendices 55

Appendix A – Final Outcome of the Study 55

Appendix B – Results of Content Analysis 67

References 87

Bibliography 90

List of Figures and Tables

Figures

Figure 1: Content Analysis Process 32

Tables

Table A-1: Metadata Management - Data Quality Technique #1........ 55

Table A-2: Data Cleansing - Data Quality Technique #2 60

Table A-3: Information Stewardship - Data Quality Technique #3........ 62

Table B-1: Metadata Management Content Analysis 67

Table B-2: Data Cleansing Content Analysis........ 76

Table B-3: Information Stewardship Content Analysis 81

Chapter I. Purpose of the Study

Brief Purpose

The purpose of this study is to examine three specific data management techniques designed to improve the value of data within a corporate data warehouse. The term “data warehouse” refers to a database of snapshots and aggregations of data throughout an enterprise (Kimball, et al., 1998). The three techniques selected for examination in this study are metadata management (Brackett, 1996; English, 1999), data cleansing (Eckerson, 2002; Kimball & Caserta, 2004), and information stewardship (Kelly, 1995; English, 1999).

The study is designed for information technology (IT) data warehousing business analysts, developers, and managers interested in improving their data warehouses. The goal of the study is to provide these professionals with a synthesized view of key concepts of three data management techniques that increase data quality within a data warehouse: metadata management, data cleansing, and information stewardship (English, 1999). According to Redman (1996), implementation of a data quality program utilizing proven data quality techniques can greatly improve the strategic value and success of a data warehouse.

The corporate data warehouse provides a single source of strategic information (Kelly, 1995) to support organizational decision-making (Kimball, et al., 1998). According to Kelly (1995), an organization’s enterprise data are strategic assets that provide competitive advantage. Market leaders maintain competitive advantage by valuing data and encouraging organizational knowledge through data analysis (Huang, et al., 1999).

Brackett (1996) argues that organizations require more information to accommodate rapidly changing business needs, yet “most organizations have large and growing quantities of disparate data” (p. 5), data that are ambiguous and unclear. Redman (1996) states that because decision-making involves risk, decisions based upon poor data increase risk for an organization. As a result, it stands to reason that a data warehouse project will have a high chance of failure if the data is of poor data quality, resulting in inaccurate organizational decisions and limited strategic information (English, 1999).

The study is designed as a literature review (Leedy & Ormrod, 2001) of materials published between 1992 and 2004 on the topics of data quality and data warehouses. Content analysis is applied to collected texts, using selective reduction on the three pre-determined key phrases “metadata management”, “data cleansing”, and “information stewardship”, allowing for identification of generalized meanings and synonyms of these three data management techniques (CSU Writing Lab, 2005). Synonyms revealed in the literature are grouped thematically by each technique and tabulated in the content analysis results (Krippendorff, 2004) (see Tables B-1, B-2 and B-3 in Appendix B).

The results of the content analysis are reorganized in the final outcome of the study, designed as a table which is divided into three rows that represent each of the data management techniques: metadata management, data cleansing, and information stewardship (see Table A-1: *Metadata Management: Data Quality Technique #1*, Table A-2: *Data Cleansing: Data Quality Technique #2*, and Table A-3: *Information Stewardship: Data Quality Technique #3* in Appendix A). For each data management technique, the list of themes and synonyms identified during content analysis is presented. Tables A-1, A-2, and A-3 are intended for use by IT data warehousing business analysts, developers, and managers as a tool to assist in the design of the metadata management, data cleansing, and information stewardship components for improvement of the value of data within a corporate data warehouse. The presentation of synonyms is intended to provide the IT data warehousing professional with clarification of data quality terminologies to reduce confusion related to the study of data management techniques. The presentation of themes is intended to provide the IT data warehousing professional with a profile of each pre-selected data quality technique and may also provide the reader with a set of evaluation criteria when assessing data quality tools.

Full Purpose

Market leaders are realizing that sources of data within an organization are strategic assets (Kelly, 1995; Loshin, 2003). An organization that values and analyzes its data can maintain competitive advantage (Huang, et al., 1999). Technology-savvy organizations are leveraging their data, defined as units of facts with specific meaning within a period of time, to create information, defined as a collection of data that has meaning to a user at a point in time (Brackett, 1994). Information drives strategic decision-making and organizational knowledge, which are the core capabilities of successful organizations (Galliers & Baets, 1998).

To make information more readily available and usable by individuals within an organization, information technology (IT) departments are implementing data warehouses (Huang, et al., 1999). A data warehouse facilitates information access by providing a centralized database for all enterprise data organized in a manner specifically for querying (Kimball, et al., 1998). The data are arranged in subject areas important to the organization and provide a static, consistent view of information, traits not found in the operational systems capturing business transactions (Kimball, et al., 1998; Loshin, 2003).

The development and maintenance of a data warehouse typically require a dedicated staff of IT professionals with particular data warehousing skills (Kimball, et al., 1998). Business analysts help to determine the subject areas and sources of data (Kimball, et al., 1998). Developers are needed to create the programs that *extract* source data, *transform* the data into usable subject areas, and *load* the data (known as Extract, Transform, & Load or ETL) into the data warehouse (Kimball & Caserta, 2004). Data access developers support the tools used by the end-users to query data, and database developers maintain the database containing all of the aggregations (stored summaries of data for faster queries) and snapshots (static views of data representing a time period) of enterprise data (Kimball, et al., 1998; Loshin, 2003).

Although a data warehouse is intended to provide valuable information for strategic decision-making, its Achilles heel is poor data quality (Brackett, 1996). According to Loshin (2003), high quality data in the data warehouse is “probably the most important success factor, because if the information is low quality, no other component of the system can be trusted" (p. 35). Further emphasizing the need for high quality data in a data warehouse, organizations require more information to accommodate rapidly changing business needs (Brackett, 1996).

Redman (2001) defines data quality as the degree to which data are useful to a specific user for a specific business need. The four dimensions of data quality are (Redman, 1996):

- **Accuracy –** The precision of a data value to the acceptable range of values.
- **Currency –** The determination of whether the data value is up-to-date.
- **Completeness –** The degree to which values are populated in a collection.
- **Consistency –** The level of overlap of entities and attributes.

This study investigates the roles certain data management techniques play in maintaining high data quality in a data warehouse. The study focuses on the three data management techniques of *metadata management*, *data cleansing*, and *information stewardship*. Metadata "include names, definitions, logical and physical data structure, data integrity, data accuracy, and other data about the organization's data resource" (Brackett, 1994, p. 451). *Metadata management* is the administration of the metadata resource. *Data cleansing* is the automated or manual process of correcting data errors (Loshin, 2003). *Information stewardship* is the responsibility of information stewards, the people who are accountable for a certain aspect of the information resource (English, 1999).

The study is designed as a literature review (Leedy & Ormrod, 2001) drawing from literature on the subjects of data quality and data warehousing. Leedy & Ormrod (2001) suggest that literature review allows a researcher to more effectively undertake a research problem by understanding existing perspectives on the topic. Materials to be reviewed are 14 articles published from 1992 to 2004. The beginning time frame of 1992 represents the year in which data quality authority Thomas C. Redman published his first book, *Data Quality Management and Technology*. Sources for the literature are University of Oregon Libraries Online, Portland State University library, and the World Wide Web.

Collected literature is subjected to content analysis, as this process is described in the CSU Writing Lab (2005). The techniques of data cleansing, metadata management, and information stewardship are used to frame the content analysis, because they provide significant value for a data quality program targeted for a data warehouse (Olson, 2003). Conceptual analysis is applied to collected texts, using selective reduction on the three pre-determined key phrases “metadata management”, “data cleansing”, and “information stewardship” (CSU Writing Lab, 2005). Analysis progresses through three general phases, described in greater detail in the Method chapter: (1) determination of concepts (*metadata management*, *data cleansing*, and *information stewardship*), (2) reading and coding of texts (discovery of selected concepts and synonyms), and (3) reporting and analyzing results (grouping of themes, creation of tables and writing conclusions) (CSU Writing Lab, 2005).

Reading and coding for the existence of the terms allows for identification of generalized meanings and synonyms of these data management techniques (CSU Writing Lab, 2005). A preliminary review of the literature reveals that there are interchangeable terminologies in use.

The results of the conceptual analysis process are formatted as a set of three tables (see Tables B-1, B-2 and B-3 in Appendix B), designed to present synonyms found in the literature. Synonyms are tabulated (Krippendorff, 2004) and listed by each technique. Themes are identified by reviewing the listings for each technique and grouping occurrences of similar and/or related concepts (Krippendorff, 2004). The themes and synonyms are aligned and presented in the final outcome of the study (see Table A-1: *Metadata Management: Data Quality Technique #1*, Table A-2: *Data Cleansing: Data Quality Technique #2*, and Table A-3: *Information Stewardship: Data Quality Technique #3* in Appendix A).

Table A-1: *Metadata Management: Data Quality Technique #1*, Table A-2: *Data Cleansing: Data Quality Technique #2,* and Table A-3: *Information Stewardship: Data Quality Technique #3* in Appendix A are the results of the content analysis. Each table represents one of the pre-selected data management techniques: metadata management, data cleansing, and information stewardship (see Table A-1: *Metadata Management: Data Quality Technique #1*, Table A-2: *Data Cleansing: Data Quality Technique #2,* and Table A-3: *Information Stewardship: Data Quality Technique #3* in Appendix A). For each data management technique, the list of themes and synonyms identified during content analysis is presented.

This outcome is intended for use by IT data warehousing business analysts, developers, and managers as a tool to assist in the design of metadata management, data cleansing, and information stewardship techniques for improvement of the value of data within a corporate data warehouse. The presentation of synonyms is intended to provide the IT data warehousing professional with an overview of the lexicon currently in use and a cross-reference of terminologies, designed to clarify the language and reduce confusion related to the study of data management techniques. Themes are intended to provide the IT data warehousing professional a rich, synthesized profile of each pre-selected data quality technique and its critical functions within the overall data quality program. Themes may also provide the reader with a set of evaluation criteria when assessing data quality tools.

Significance of the Study

According to Huang, et al. (1999), a corporate data warehouse provides tools for an organization to cultivate organizational knowledge and gain competitive advantage. Maintenance of high quality data in a data warehouse is a key success factor for IT data warehousing professionals (Loshin, 2003). Inaccurate data significantly weakens the strategic value of a data warehouse (Faden, 2000). Consequently, poor data quality can result in an ineffective data warehouse and has been found to be a leading cause for the failure of a data warehouse project (Hudicka, 2003).

This study seeks to understand literature that addresses the data management techniques that data warehousing IT professionals can employ to counteract the negative outcomes of poor data quality in a data warehouse. A review of the literature in this study reveals that in most cases collected books and articles fit into one of three categories: 1) a general guide for enterprise data quality (e.g. Brackett, 1994, 2000; Huang, et al., 1999; Olson, 2003; Redman, 1992, 1996, 2001; St. Clair, 1997); 2) a general guide for data warehousing (e.g. Kelly, 1994; Kimball, et al., 1998; Kimball & Caserta, 2004; Loshin, 2003); or, 3) a data quality guide aimed at data warehousing (e.g. Brackett, 1996; Eckerson, 2002; English, 1999).

The information contained in literature collected for review in this study, unfortunately, is not presented in a manner that clearly explains the components of a data quality program suitable for a data warehouse project. This study attempts to filter out techniques not applicable for a data warehouse project and clearly delineate data quality program components, thereby providing a more valuable tool for IT data warehousing professionals exploring a data quality initiative.

Limitations to the Research

The information gathered for research is drawn from literature on the subjects of data quality and data warehousing. The majority of content is derived from authoritative books on the subjects, with a start date of 1992. The beginning time frame of 1992 was chosen to allow inclusion of data quality authority Thomas C. Redman’s first book, *Data Quality Management and Technology*, published in 1992.

Articles dating from 1995 to 2005 are also analyzed to provide current statistics for information represented in the *Significance of Study* and *Problem Area* sections, to obtain up-to-date information about software vendors in the data quality market, and to validate the currency of data quality techniques presented in the books under review. The study is not intended to be an exhaustive analysis of all relevant literature for data quality and data warehousing, but a sampling of over 40 authoritative books and articles, selected from within this time frame.

On the topic of data quality, books published by recognized authors on the subject are given precedence. For example, literature includes three books by Thomas C. Redman, a prolific author in the field of data quality and a former manager of data quality at AT&T Bell Labs (Redman, 1992, 1996, 2001), three books by Michael H. Brackett, a well-known authority on data management and data quality (Brackett, 1994, 1996, 2000), and literature from Larry P. English, a leading expert in the field of information quality and founder of the International Association of Information and Data Quality (English, 1999).

On the topic of data warehousing, focus is on books published by recognized authors. In particular, literature includes books co-authored by Ralph Kimball, the founder of data warehousing and an active data warehousing visionary (Kimball, et al. 1998; Kimball & Caserta, 2004).

By limiting the context of the content analysis (Krippendorff, 2004) to a data warehouse, the data quality component of *data processing improvements* is excluded. Data processing improvements entail tracking data through an operational process and making adjustments to processes for higher data quality (Redman, 1996). While there may be minimal benefits of data processing improvement in a data warehousing ETL process, generally the technique is more suitable to an operational system. The techniques of data cleansing, metadata management, and information stewardship are included in the content analysis, because they do provide significant value for a data quality program targeted for a data warehouse. A broader, more comprehensive study of data quality for data warehouses would include data processing improvements.

The study utilizes selective reduction on the phrases “metadata management”, “data cleansing”, and “information stewardship” to focus the research on specific categories (CSU Writing Lab, 2005). The process of selective reduction allows for generalized meanings and synonyms of these three phases, which is useful in this case, because the literature has shown that there are interchangeable terminologies in use for data quality techniques. For example, the term “information steward” is also referred to as the “information product manager” (Huang, et al., 1999, p. 26), “data custodian, data steward, or data trustee” (Redman, 1996, p. 51). Table A-1: *Metadata Management: Data Quality Technique #1*, Table A-2: *Data Cleansing: Data Quality Technique #2*, and Table A-3: *Information Stewardship: Data Quality Technique #3* in Appendix A contain synonyms for the terms for easy reference by IT data warehousing professionals.

In addition to synonyms for each data quality technique, Tables A-1, A-2, and A-3 include identification of themes grouped within each technique. The themes are identified by the researcher and reflect personal understanding and interpretation of the concepts.

Purely technical implementations of a theme such as the inclusion of primary key references in a data definition (Brackett, 2000) are filtered out or incorporated into a more broad-based explanation, tied directly to the focus of the study. The final outcome of this study is meant to provide IT data warehousing professionals with a clear, synthesized, and organized view of the data quality techniques of metadata management, data cleansing, and information stewardship, not a technical design document. Readers interested in the technical details or in the discovery of additional data quality themes are encouraged to reference the *Bibliography* for further reading.

Definitions

Aggregate. “Aggregates are stored summaries built primarily to improve query performance” (Kimball, et al., 1998, p. 211).

Content analysis. “Content analysis is a research tool used to determine the presence of certain words or concepts within texts or sets of texts” (CSU Writing Lab, 2005).

Corporate data warehouse. “[A] *subject-oriented* corporate database which addresses the problem of having multiple data models implemented on multiple platforms and architectures in the enterprise” (Kelly, 1995, p. 6).

Data. “The individual facts with a specific meaning at a point in time or for a period of time” (Brackett, 1994, p. 421).

Data cleansing. “Data cleansing is the process of finding errors in data and either automatically or manually correcting the errors” (Loshin, 2003, p. 236).

Data processing improvements. The tracking of data through an operational process and making adjustments to processes for higher data quality (Redman, 1996).

Data profiling. “[Data profiling] employs analytical methods for looking at data for the purpose of developing a thorough understanding of the content, structure, and quality of the data” (Olson, 2003, p. 20)

Data quality. “The degree to which data meet the specific needs of specific customers” (Redman, 2001, p. 223). The dimensions of data quality are “accuracy, currency, completeness, and consistency” (Redman, 1996, p. 254).

Data quality program. An enterprise-level data quality initiative with “clear business direction, objectives, and goals”, “[m]anagement infrastructure that properly assigns responsibilities for data”, “[a]n operational plan for improvement”, and “[p]rogram administration” (Redman, 1996, pp. 18-19).

Data warehouse. A database of snapshots and aggregations of data throughout an enterprise to be used for querying and decision-making (Kimball, et al., 1998).

Disparate data. “Data that are essentially not alike, or are distinctly different in kind, quality, or character” (Brackett, 1994, p. 441).

ETL (Extract, Transform, & Load). “[The] sequence of applications that extract data sets from the various sources, bring them to the data staging area, apply a sequence of processes to prepare the data for migration into the data warehouse, and then actual loading the data” (Loshin, 2003, p. 246).

Information. “A collection of data that is relevant to the recipient at a point in time. It must be meaningful and useful to the recipient at a specific time for a specific purpose” (Brackett, 1994, p. 446).

Information quality. “The degree to which data are transformed into information to resolve uncertainty or meet a need” (Brackett, 1994, p. 447).

Information steward. “The role of people with respect to their accountability for the integrity of some part of the information resource” (English, 1999, p. 479).

Literature review. “Research proposals and research reports typically have a section (in the case of a thesis or dissertation, an entire chapter) that reviews the related literature.

The review describes theoretical perspectives and previous research findings related to the problem at hand" (Leedy & Ormrod, 2001, p. 70).

Metadata. "Data about the data. They include names, definitions, logical and physical data structure, data integrity, data accuracy, and other data about the organization's data resource" (Brackett, 1994, p. 451).

Operational system. "An operational system of record whose function is to capture the transactions of the business" (Kimball, et al. 1998, p. 14).

Selective reduction. "The central idea of content analysis. Text is reduced to categories consisting of a word, set of words or phrases, on which the researcher can focus. Specific words or patterns are indicative of the research question and determine levels of analysis and generalization" (CSU Writing Lab, 2005).

Snapshot. A static view of data representing a period of time (Kimball, et al., 1998).

Strategic asset. A strategic asset "can be used to provide benefits to the company, is controlled by the organization, and is the result of previous transactions" (Loshin, 2003, p. 12).

Strategic decision-making. "[L]ong-term decisions that affect the health of the enterprise or its important organization over the long term" (Redman, 1996, p. 44).

Strategic information. "Data [that are] required as a catalyst for creative thinking, or for accurate analysis, or in acquiring or allocating resources, or in assessing the impact of events, or in evaluating alternative actions, or in establishing general principles" (Kelly, 1995, p. 23).

Problem Area

Competitive organizations are now implementing data warehouses to cultivate organizational knowledge, gain insight into business performance trends, better understand and serve their customers, and solve business problems (Huang, et al., 1999). The growth of the data warehousing industry is one measure of the value that organizations are placing on the strategic importance of data warehouses. According to the On Line Analytical Processing (OLAP) Report, the data warehousing industry quadrupled in market size from $1 billion in 1996 to $4.3 billion in 2004 (Pendse, 2005).

Despite this incredible growth, Gartner predicts that 50% of data warehouse projects through 2007 will have limited success or will fail (Beal, 2005). A significant cause of data warehouse project failures is poor data quality (Hudicka, 2003). Brackett (1996) has found that most companies experience a problem with growing volumes of disparate data, defined as data that is ambiguous or unclear.

According to Faden (2000), the ability of an organization to make accurate strategic decisions is greatly weakened when the data warehouse contains inaccurate data. When decision-makers retrieve low quality, disparate data from the data warehouse, they 1) don’t trust the data; 2) proceed cautiously with results; and, 3) seek confirmatory data, all of which add time, confusion, and risk to the decision-making process (Redman, 2001). Consequently, a data warehouse project will have a high chance of failure if it facilitates inaccurate organizational decisions and limited strategic information (English, 1999).

Furthermore, poor data quality costs U.S. businesses over $600 billion annually (Eckerson, 2002) and consumes approximately 10% of the typical organization’s revenue (Redman, 2001). Recent high profile news stories of data quality issues such as the 2000 U.S. Presidential election, incorrect pricing on Amazon’s website, and delayed Christmas toy shipments from Toys ‘R Us illustrate the crippling effects of poor data quality (Redman, 2001).

This study seeks to understand literature that addresses the data management techniques that data warehousing IT professionals can employ to counteract the negative outcomes of poor data quality in a data warehouse. A review of the literature in this study has uncovered that in most cases books and articles fit into one of three categories:

- *A general guide for enterprise data quality* (e.g. Brackett, 1994, 2000; Huang, et al., 1999; Olson, 2003; Redman, 1992, 1996, 2001; St. Clair, 1997). Methods described in this category of sources are outlined for the implementation of a data quality program to champion data quality improvements throughout an enterprise. The data quality program typically contains the techniques of data profiling and cleansing, metadata management, information stewardship, and data processing improvements.

- *A general guide for data warehousing* (e.g. Kelly, 1994; Kimball, et al., 1998; Kimball & Caserta, 2004; Loshin, 2003). Authors in this category of sources explain the fundamentals of data warehousing and the resources necessary to implement a data warehouse project. Often the material is technical in nature and is organized by the data warehousing layers of data access, ETL, and database modeling. Data quality topics focus on data cleansing in the ETL layer and metadata management.
- *A data quality guide aimed at data warehousing* (e.g. Brackett, 1996; Eckerson, 2002; English, 1999). A data quality program is outlined in these sources with the same data quality components of data profiling and cleansing, metadata management, information stewardship, and data processing improvements found in the general enterprise data quality guides. Certain guidelines focus on data warehousing concepts, primarily in the area of ETL and database modeling.

While the information contained in literature collected for review in this study explains data quality techniques, the components of a data quality program suitable for a data warehouse project are ambiguous. The materials in the first and third category intertwine concepts and include some techniques more appropriate for the operational systems, but not for the data warehouse. The materials in the second category provide limited context for topics on data quality and bury data quality concepts in technical jargon. This study attempts to filter out techniques not applicable for a data warehouse project and also to clearly delineate data quality program components, thereby providing a more valuable tool for IT data warehousing professionals who are exploring a data quality initiative.

Chapter II. Review of References

The *Review of References* contains a summary of key references used to establish a research methodology and conduct the data analysis. The data analysis literature is presented in three categories: 1) general guides for enterprise data quality, 2) general guides for data warehousing, and 3) data quality guides aimed at data warehouses.

The *Review of References* takes the form of an annotated bibliography. Each review provides a summary of the relevant content, a reason for selection in the study, and an explanation of its contribution to the study.

Key literature supporting the research method

CSU Writing Lab (Colorado State University). (2005). “Conducting Content Analysis”. *Writing Guides*. Retrieved September 26, 2005 from http://writing.colostate.edu/guides/research/content/index.cfm.

This website provides a practical guide to carrying out a content analysis for a research project. It contains eight consecutive steps for conducting conceptual analysis. The University of Oregon AIM Program recommends The CSU Writing Guide as a resource for academic and research writing guidelines. This study employs the eight steps of conceptual analysis, detailed in the *Method* chapter. In addition, the approach to content analysis recommended by the CSU Writing Lab defines the framework for the *Analysis of Data* chapter.

Key literature concerning enterprise data quality

Brackett, Michael H. (2000). *Data Resource Quality: Turning Bad Habits into Good Practices*. Upper Saddle River, NJ: Addison-Wesley.

This book is a practical guide to data quality techniques for the enterprise data resource. It opens with a general discussion of the business importance of data quality. For each of the outlined data quality strategies, the strategy is explained simply and its business value is highlighted. Of particular importance to this study is the content related to metadata management and data stewardship, because these topics constitute a significant portion of the book.

This resource is selected for the study because its author, Michael H. Brackett, is a frequently cited author on the subject of enterprise data architecture with an emphasis on data quality. Data quality techniques in the book are presently clearly and strike a good balance between technical and non-technical verbiage. The metadata management and data stewardship topics factor into the *Analysis of Data* and *Conclusions* chapters.

Brackett, Michael H. (1994). *Data Sharing*. New York: John Wiley & Sons, Inc.

This book discusses processes an enterprise should have in place to ensure that data across an enterprise is usable. The author explains how disparate data can reduce the value of enterprise data and how formal data documentation can control disparate data.

In addition, Brackett recommends a facilitated approach to data documentation to ensure consensus across the enterprise.

This resource is written by the well-known data quality author Michael H. Brackett. Brackett presents the topics in simple language, highlighting key concepts. The book is a source for the *Analysis of Data* and *Conclusions* chapters as a rich resource for metadata management techniques in its discussion of data documentation and some insight into information stewardship in its sections on a facilitated approach to data documentation.

Huang, Kuan-Tse, Lee, Yang W., & Wang, Richard Y. (1999). *Quality Information and Knowledge*. Upper Saddle River, NJ: Prentice-Hall.

This book stresses the importance of information quality as a critical component for organizational knowledge. The authors explain how quality information is an asset for an organization. The book describes how to set up an information quality program to maintain high quality information.

The authors were a professor, associate professor, and Ph.D. recipient from the Massachusetts Institute of Technology (MIT), and two were leaders of the Total Data Quality Management program at the school at the time of publication. In addition, the book contains an extensive list of information quality and knowledge management references and bibliography. The discussion of an information quality program factors into the *Analysis of Data* and *Conclusions* chapters for its information stewardship components. The resource is also referenced in the *Purpose of the Study* to help build the case for the *Significance of the Study* and *Problem Area*.

Olson, Jack E. (2003). *Data Quality: The Accuracy Dimension*. San Francisco: Morgan Kaufmann Publishers.

This book is a mix of non-technical and technical methods for data quality. It contains three sections: a definition of data accuracy; a guide for developing a data quality assurance program; and, an outline of data profiling technology features. The data profiling features include a data profiling repository and data cleansing.

Olson has been working in the data management field for over 35 years. After a long career at major information technology firms, he is currently chief technology officer at Evoke Software, where he originated the idea of data profiling.

This resource factors into the *Analysis of Data* and *Conclusions* chapters. The data profiling repository is included in the metadata management component, the data quality assurance program in information stewardship, and the data cleansing feature of data profiling technology in data cleansing.

Redman, Thomas C. (1996). *Data Quality for the Information Age.* Boston: Artech House.

This book is a mix of non-technical and technical guidelines for the implementation of a data quality initiative that encompasses a formal data quality policy, measurement and control, and process improvements. The reference also contains case studies of data quality programs at AT&T and Telstra. While a significant portion of the book discusses information processing and reengineering strategies not applicable to a data warehouse, it includes valuable insights into information stewardship, metadata management, and data cleansing.

Redman is a frequently cited author in the field of data quality and a former manager of data quality at AT&T Bell Labs. In this book, he expands upon the data quality themes he introduced in his 1992 book *Data Quality Management and Technology*. This reference is used in *Analysis of Data* and *Conclusions* chapters. The data quality policy and program concepts factor into information stewardship, data measurement concepts into metadata management, and data error correction concepts into data cleansing.

Redman, Thomas C. (1992). *Data Quality Management and Technology*. New York: Bantam Books.

This book introduces the notion that quality of data is as important as the quality of business processes. Redman introduces a definition of data quality and outlines the dimensions of data quality. He explains what the components of a successful data quality program should contain.

This resource provides a groundbreaking definition of data quality from the prolific data quality author and former manager of data quality at AT&T Labs, Thomas C. Redman. The concepts in the book are applied to the study throughout the *Purpose of the Study* and are a reference in the *Analysis of Data* and *Conclusions* chapters. Information stewardship themes are found in the data quality program discussion, metadata management themes in the data measurement systems discussion, and data cleansing in the database cleanup discussion.

Redman, Thomas C. (2001). *Data Quality: The Field Guide.* Boston: Digital Press.

This book provides an introduction to data quality for information technology professionals who are unfamiliar with its components. It also provides anecdotes, case studies, and statistics regarding data quality that can be used to build the case for implementing a data quality program.

This resource is written by Thomas C. Redman and summarizes concepts he has written about in earlier works on data quality. In this book, he simplifies data quality ideas and champions the need for data quality initiatives. This study draws from the case studies and statistics presented in the book to include in the *Problem Area* of the *Purpose of the Study*. In addition, the resource is utilized in *Analysis of Data* and *Conclusions* chapters for its information stewardship, metadata management, and data cleansing themes.

Key literature concerning data warehousing

Kelly, Sean. (1995). *Data Warehousing: The Route to Mass Customization*. New York: John Wiley & Sons.

This book is a management-oriented description of data warehousing. It explains the history, business drivers, and components of a data warehouse. This book gives a good high-level overview of the resources needed to manage a successful data warehouse, including a brief description of where data quality fits into the system.

Kelly has written extensively on the data warehousing industry. This book is his first major work on the topic and succinctly outlines the business and technical resources involved in a data warehousing project. This resource is applied in the study to identify the high level components of a data warehouse and to better understand what data quality techniques apply to data warehousing. It is specifically used in the *Analysis of Data* and *Conclusions* chapters for its references to data quality in a data warehouse.

Kimball, Ralph, & Caserta, Joe. (2004). *The Data Warehouse ETL Toolkit*. New York: Wiley Computer Publishing.

This book focuses on the technical aspects of the Extract, Transform, & Load (ETL) phase of a data warehouse. It provides detailed guidelines for developing an effective ETL process that efficiently loads high quality data in a presentation-ready format into the data warehouse from source systems. There are chapters specifically dedicated to the technical implementation of data quality for a data warehouse.

This resource is co-authored by the leading visionary in the data warehousing industry since nearly its inception, Ralph Kimball. This book is a more detailed look at the ETL process, the phase where data quality can be most suitably enforced in a data warehouse. The reference is used in the *Analysis of Data* and *Conclusions* chapters, because it addresses in-depth metadata management and data cleansing for a data warehouse.

Kimball, Ralph, Reeves, Laura, Ross, Margy, & Thornhwaite, Warren. (1998). *The Data Warehouse Lifecycle Toolkit*. New York: Wiley Computer Publishing.

This reference contains a comprehensive guide to the technical development of a data warehouse. It covers the full spectrum of source data analysis, ETL, and data access. Data cleansing and other data quality techniques are referenced throughout the book.

This seminal resource for the development of data warehouse is co-authored by data warehousing pioneer Ralph Kimball. The book is often referred to as the “data warehousing bible” in the data warehousing industry. The book is referenced in the study as a resource for industry-standard terminology in the *Full Purpose*. It is also a source in *Analysis of Data* and *Conclusions* chapters for its inclusion of data quality techniques for data warehouses.

Loshin, David. (2003). *Business Intelligence*. San Francisco: Morgan Kaufmann Publishers.

This book is geared toward managers who want to launch a data warehousing project. It contains all of the key high-level concepts of data warehousing and metrics for ensuring success of the project. The author identifies data quality as a key success factor for a data warehousing project and includes a section emphasizing data quality and its role in a data warehouse.

The author has written multiple information technology books and this resource is part of the *The Savvy Manager's Guide* series, books geared toward managers researching certain technology-related projects. The resource is used in the *Full Purpose* to build the case for the value of data quality in a data warehouse. In addition, it is included as a reference in the *Analysis of Data* and *Conclusions* chapters for its insights into metadata, data cleansing, and data ownership (an information stewardship synonym).

Key literature concerning data quality for data warehouses

Brackett, Michael H. (1996). *The Data Warehouse Challenge: Taming Data Chaos*. New York: Wiley Computer Publishing.

This book applies the data quality themes the author has formulated in earlier works to the development of a data warehouse. As with Brackett's other books, the author makes the case that disparate data must be controlled. Brackett highlights the special role the data warehouse plays as a central data repository in helping to maintain enterprise data quality. Many of the specific data quality techniques are presented similarly to his other works with an emphasis on data warehousing where applicable.

This resource is one of many books authored by Michael H. Brackett, a well-known data quality thinker. Its focus on data warehousing provides additional value for the study. The resource is referenced in the *Purpose of the Study* in the *Problem Area* as an explanation for disparate data. It also is included in the *Analysis of Data* and *Conclusions* chapters, because it covers all three data quality techniques extensively.

Eckerson, Wayne W. (2002). "Data Quality and the Bottom Line". *TDWI [Online]*. Retrieved September 13, 2005 from http://www.tdwi.org/research/display.aspx?ID=6028.

This thirty-two page report discusses the importance of data quality in a data warehouse. It presents industry statistics on the problems of poor data quality, a plan for creating a data quality program, and a primer on how to select a data cleansing software tool.

The author of the report is a director at The Data Warehousing Institute (TDWI), an educational organization for the data warehousing industry. The resource provides valuable statistics for the Problem Area section of the *Purpose of the Study*, and it includes information about data cleansing tools and information stewardship via a data quality program for the *Analysis of Data* and *Conclusions* chapters.

English, Larry P. (1999). *Improving Data Warehouse and Business Information Quality*. New York: Wiley Computer Publishing.

The technical book provides specific measures for ensuring information quality in a data warehouse. The first section defines information quality in detail. The second section explains processes that ensure information quality. The third section discusses how to keep information quality intact through information stewardship programs.

This resource is written by Larry P. English, a leading consultant on data quality. English is also founder of the International Association of Information and Data Quality. This resource provides the study with the term "information stewardship". It is also a key source in the *Purpose of the Study* and in *Analysis of Data* and *Conclusions* chapters with references to metadata management, data cleansing and information stewardship.

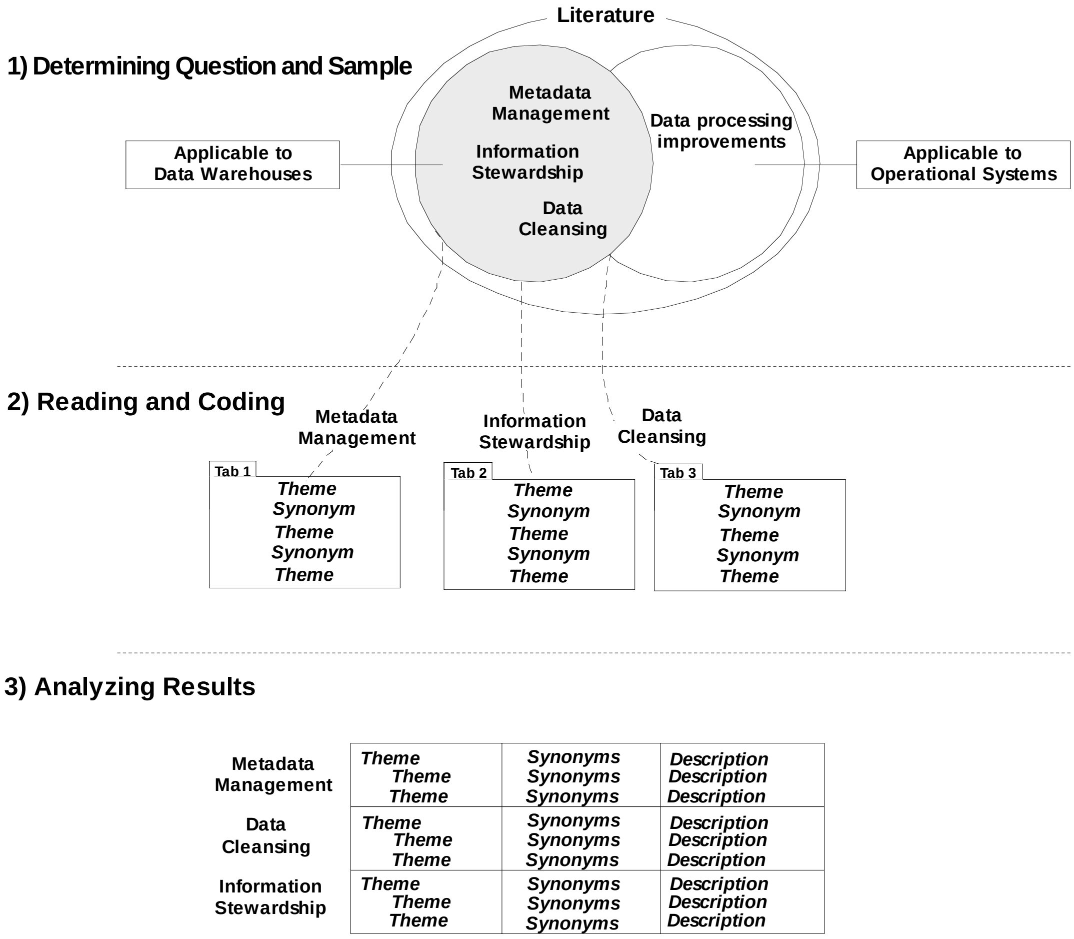
Chapter III. Method

The study is designed primarily as a Literature Review (Leedy & Ormrod, 2001). Leedy & Ormrod (2001) propose literature review, because “the more you know about investigations and perspectives related to your topic, the more effectively you can tackle your own research problem” (p. 71).

Collected literature is subjected to content analysis adhering to the process described in the CSU Writing Lab (2005):

1) Data collection: Identify a research question and choose a sample of literature.
2) Data analysis: Code text into manageable content categories.
3) Data presentation: Analyze the results.

What follows is an outline of the methodology plan used in this study. Figure 1 is a graphic representation of this process.



Metadata Management	*Theme* *Theme* *Theme*	*Synonyms* *Synonyms* *Synonyms*	*Description* *Description* *Description*
Data Cleansing	*Theme* *Theme* *Theme*	*Synonyms* *Synonyms* *Synonyms*	*Description* *Description* *Description*
Information Stewardship	*Theme* *Theme* *Theme*	*Synonyms* *Synonyms* *Synonyms*	*Description* *Description* *Description*

Figure 1: Content Analysis Process

Research question

The research question that guides this study is: “How can information technology (IT) data warehousing business analysts, developers, and managers improve the quality of data in their data warehouses?”

Data collection – Choosing a sample of literature

Literature published from 1992 to 2005 is sampled, referencing over 40 authoritative books and articles on the subjects of data quality and data warehousing. The beginning time frame of 1992 represents the year data quality authority Thomas C. Redman published his first book, *Data Quality Management and Technology*.

This study is based on literature gathered from the following sources:

- **University of Oregon Libraries Online**
    - **EBSCOhost Research Databases and Lexis/Nexis Academic.** Provides full-text online archives of periodicals on the subjects of data quality and data warehousing. Krippendorff (2004) recommends content providers such as Lexis/Nexis, because they verify that literature originates from trustworthy publications.
    - **University of Oregon Libraries Catalog.** Provides books from the University of Oregon library on the subjects of data quality and data warehousing.
    - **The Orbis Cascade Alliance Union Catalog.** Provides books from libraries of participating universities on the subjects of data quality and data warehousing.
- **Portland State University.** Provides books from the Portland State University library on the subjects of data quality and data warehousing.

- **World Wide Web.** Provides periodicals and articles retrieved from search engines. Only articles retrieved from credible online IT publications, software vendors, and IT associations are permissible to ensure validity of the retrieved texts (Krippendorff, 2004). Materials from software vendors are strictly used to determine features of their data quality products and are not treated as unbiased resources for data quality and data warehousing insights.

Data analysis - Coding text into manageable content categories

Conceptual analysis is applied to collected texts, using the eight steps for conducting conceptual analysis as described in the CSU Writing Lab (2005).

Step 1: Decide the level of analysis - The technique of selective reduction is used to determine which phrases make up a concept. The CSU Writing Lab (2005) defines selective reduction as:

> [T]he central idea of content analysis. Text is reduced to categories consisting of a word, set of words or phrases, on which the researcher can focus. Specific words or patterns are indicative of the research question and determine levels of analysis and generalization.

Step 2: Decide how many concepts to code - The three pre-determined key phrases selected as the coding concepts in this study are “metadata management” (Brackett, 1994), “data cleansing” (Loshin, 2003), and “information stewardship” (English, 1999).

These phrases are used to represent the three data quality techniques applicable to data warehouses.

Step 3: Decide whether to code for existence or frequency - Reading and coding for the existence of the terms allows for identification of generalized meanings and synonyms (CSU Writing Lab, 2005) of these data management techniques.

Step 4: Decide how to distinguish among concepts - A preliminary review of the literature reveals that there are interchangeable terminologies in use for the three terms, such as “data description” (Brackett, 1996, p. 69) for metadata, “data-quality screen” (Kimball & Caserta, 2004, p. 114) for data cleansing, and “information product manager” (Huang, et al., 1999, p. 26) for information steward. Synonyms are noted through contextual reading of the materials.

Step 5: Develop rules for coding - Synonyms for each of the three concepts are tabulated so that iterative coding for the key concepts and synonyms can be performed. For example, when a synonym for “metadata management” is discovered, that synonym is added to the list of key phrases to be coded in the subsequent literature and during a second pass of the previously coded literature.

Step 6: Decide what to do with “irrelevant" information - By limiting the context of the content analysis (Krippendorff, 2004) to a data warehouse, themes related to the data quality component of data processing improvements, a technique suited for an operational system, is excluded. In addition, themes related to a concept that are not relevant to a data warehouse or are purely technical in nature are also discarded.

Step 7: Code the texts - Coding is performed manually by reading through the text and writing down occurrences of the concept. A Microsoft Excel worksheet with three spreadsheets, one for each concept, is the repository for recording of occurrences. Each tab has four columns with the headings “Description”, “Reference”, “Page Number”, and “Theme/Synonym”. For each theme or synonym of a concept discovered in the literature, a record is created in the appropriate spreadsheet. The record contains the description, reference material, reference page number (if applicable), and an indicator for whether the record is a theme or synonym (valid values are “theme” and “synonym”).

Data presentation – Analyzing results

Step 8: Analyze the results - *Chapter IV. Analysis of Data* presents the results of the coding exercise. The contents of the spreadsheet used for recording themes and synonyms are presented as three tables (see Table B-1: *Metadata Management Analysis*, Table B-2: *Data Cleansing Analysis* and Table B-3: *Information Stewardship Analysis* in Appendix B), one for each concept, with headings identical to the spreadsheet used for coding: “Description”, “Reference”, “Page Number”, and “Theme/Synonym”.

The final outcome of the study, Table A-1: *Metadata Management: Data Quality Technique #1*, Table A-2: *Data Cleansing: Data Quality Technique #2*, and Table A-3: *Information Stewardship: Data Quality Technique #3* in Appendix A are used to frame discussion in *Chapter V. Conclusions.* The *Conclusions* chapter presents what the reported results and outcomes of the study mean in relation to the purpose of the study and for the intended audience.

Table A-1: *Metadata Management: Data Quality Technique #1*, Table A-2: *Data Cleansing: Data Quality Technique #2*, and Table A-3: *Information Stewardship: Data Quality Technique #3* are located in Appendix A. Each table represents a data management technique: metadata management, data cleansing, and information stewardship. For each data management technique, the list of synonyms identified during content analysis, along with subsequent grouping of themes, is presented. The presentation format is intended to provide the IT data warehousing professional with: (1) a way to decipher data quality terminology; and (2) a profile of each of the three pre-selected data quality techniques.

Chapter IV. Analysis of Data

This chapter details the process of coding the raw data and organizing the data for presentation. This process corresponds to Step 7 (Code the texts) and Step 8 (Analyze the results) of a conceptual analysis as described in the CSU Writing Lab (2005) and as applied in this study. The overall goal of the content analysis process is to more fully understand data quality themes. *Appendix B – Results of Content Analysis* contains the final results of the coding exercise.

Code the texts

Key literature concerning enterprise data quality, data quality for data warehouses, and data quality for data warehouses from the *Review of References* chapter form the data set for coding. The 14 references are listed in the order in which they are coded:

Key literature concerning enterprise data quality

1) **Brackett, Michael H.** (1994). *Data Sharing*. New York: John Wiley & Sons, Inc.
2) **Brackett, Michael H.** (2000). *Data Resource Quality: Turning Bad Habits into Good Practices*. Upper Saddle River, NJ: Addison-Wesley.
3) **Huang, Kuan-Tse, Lee, Yang W., & Wang, Richard Y.** (1999). *Quality Information and Knowledge*. Upper Saddle River, NJ: Prentice-Hall.
4) **Olson, Jack E.** (2003). *Data Quality: The Accuracy Dimension*. San Francisco: Morgan Kaufmann Publishers.

5) **Redman, Thomas C.** (1992). *Data Quality Management and Technology*. New York: Bantam Books.

6) **Redman, Thomas C.** (1996). *Data Quality for the Information Age.* Boston: Artech House.

7) **Redman, Thomas C.** (2001). *Data Quality: The Field Guide*. Boston: Digital Press.

Key literature concerning data warehousing

8) **Kelly, Sean.** (1995). *Data Warehousing: The Route to Mass Customization*. New York: John Wiley & Sons.

9) **Kimball, Ralph, Reeves, Laura, Ross, Margy, & Thornhwaite, Warren.** (1998). *The Data Warehouse Lifecycle Toolkit*. New York: Wiley Computer Publishing.

10) **Kimball, Ralph, & Caserta, Joe.** (2004). *The Data Warehouse ETL Toolkit*. New York: Wiley Computer Publishing.

11) **Loshin, David.** (2003). *Business Intelligence*. San Francisco: Morgan Kaufmann Publishers.

Key literature concerning data quality for data warehouses

12) **Brackett, Michael H.** (1996). *The Data Warehouse Challenge: Taming Data Chaos*. New York: Wiley Computer Publishing.

13) **Eckerson, Wayne W.** (2002). "Data Quality and the Bottom Line". *TDWI [Online]*. Retrieved September 13, 2005 from http://www.tdwi.org/research/display.aspx?ID=6028.

14) **English, Larry P.** (1999). *Improving Data Warehouse and Business Information Quality*. New York: Wiley Computer Publishing.

The references within a topic are ordered by the author and the year of publication. For example, within the key literature concerning enterprise data quality, Redman's 1992 publication is coded before the 1996 and 2001 books. The rationale is that the author builds upon earlier works in later publications; therefore, the earliest work provides the conceptual groundwork for later releases.

The order in which the topics are coded is also taken into consideration. The data quality books are coded first to fully understand data quality themes. The data quality for data warehousing books are coded second with the reasoning that more data quality themes will emerge in these books than in the data warehousing books. The data warehousing books are coded third. The rationale is that data quality themes are less prevalent, since these books primarily focus on data warehousing themes, not data quality.

The coding process follows five steps:

Step 1: Document themes - Record each theme in a Microsoft Excel spreadsheet dedicated to a data quality concept with columns of "Description", "Reference", and "Page Number".

- **Definition** - The theme and the relevant citation in the format of theme phrase, a dash (-), then the citation in quotes. For example, a citation found in English (1999) pertaining to the data cleansing step of analyzing data defect types is recorded as:

> analyze data defect types - "This step analyzes the patterns of data errors for input to process improvements."

- **Reference** - A code indicating the reference for the theme. The coding scheme can be found in the Reference Key at the beginning of each table in *Appendix B – Results of Content Analysis*. In the example above, the reference code is "EN99", representing English (1999).
- **Page Number** - The page number for the citation.

A challenge in this step is the determination of what concepts of a theme to filter out as too technical, because most citations have technical aspects. This determination is necessary in order to meet a limitation noted in the *Limitations of the Research* section of the *Full Purpose* chapter. The guideline is that concepts related to the actual physical or programmatic execution of the theme are excluded. For example, the *analyze data defect types* theme cited in English (1999) includes six specific sections of an output report of data defects, such as frequency, relative cost, and impact of the data defect type (p. 266). This level of technical detail is excluded from the citations.

Step 2: Make a second pass - Once all references have been coded, perform a second pass through the literature. The rationale is based on the assumption that a complete list of themes and synonyms are documented in the first pass, so a second pass will ensure that no synonyms for themes have been overlooked.

Step 3: Group common themes together - Cluster themes into categories with similar meanings (Krippendorff, 2004). During this step, the researcher adheres to the following process:

1) Reorganize rows in the spreadsheet to group theme phrases together. For example, all citations related to data cleansing steps are ordered together.
2) Insert a row between logical groupings and add a notation of the high-level theme group. For example, a row above the *data cleansing steps* theme group is inserted with the phrase “Data Cleansing Steps”.
3) Group similar theme phrase rows together within a theme group based upon citations with very similar meanings. For example, under the *data cleansing steps* theme group, the two citations referring to the steps *analyze data defect types* are ordered together.

Step 4: Determine themes and synonyms - For citations with very similar meanings, determine which theme phrase is the theme and mark all of the other citations as synonyms. A column is added to the end of each spreadsheet called “Theme/Synonym” to record whether a row is a theme or a synonym. The determination is made through a two-step decision tree (Krippendorff, 2004):

1) Within a grouping of similar theme phrases, determine the theme by frequency of the theme phrase within a cluster of theme phrases. For example, of the seven

themes related to the definition of data profiling, two use the explicit term "data profiling" while the other five are distinctly different terms like "data auditing", "data exploration prototype", etc. Therefore, choose "data profiling" as the theme and consider the other theme phrases as synonyms; and

2) If a theme cannot be singled out using the rule above, choose the theme from the most recent data quality reference, excluding Eckerson (2002) (this reference is excluded because it is a relatively brief report, not a text book like the other references). For example, choose "data refining" as the theme over "information refining", because "data refining" is cited from a 1996 publication while "information refining" is cited from a 1992 publication.

Order the selected theme within a group of theme phrases as the first entry within the group and mark as "Theme" in the "Theme/Synonym" column. Mark all other theme phrases in the batch as "Synonym" and order underneath the "Theme".

Step 5: Organize theme groups - Organize theme groups further by adding additional sub-headings to theme groups and ordering by highest level themes to lowest-level sub-themes using a tree organization (Krippendorff, 2004). For example, organize the *data cleansing steps* themes in the correct order that the steps should be executed and give each step a sub-heading.

Present the results

The final outcome of the analysis of the results is presented in a three-column table with column headings of “Theme”, “Synonyms”, and “Description”. *Appendix A – Final Outcome of the Study* contains the final presentation of the content analysis results. Each data quality technique is represented in a different table with the themes sourced from the corresponding coding exercise spreadsheet. Below is the explanation of the content in each of the three columns:

- **Column 1: Theme** - The theme phrase from the coding exercise spreadsheet that is marked as a “Theme”. Sub-themes are indented in a manner that represents the tree organization determined in Step 5 (Organize theme groups) of the coding exercise.
- **Column 2: Synonyms** - A semicolon-delimited list of the synonyms for a theme. These are the theme phrases marked as a “Synonym” in the coding exercise spreadsheet.
- **Column 3: Description -** A synthesized description using the citations for the theme and its synonyms.

Chapter V. Conclusions

To meet the goals of this study, Appendix A – *Final Outcome of the Study, Three Data Quality Techniques: Themes*, *Synonyms & Definitions* is provided as a tool for use by IT data warehousing professionals interested in improving the data quality of their data warehouses. This table presents themes, synonyms and definitions of concepts related to three data quality techniques: metadata management, data cleansing, and information stewardship. The goals of the study can be summarized as:

- Present a clarification of terminology to reduce confusion;
- Present a profile of each data quality technique to assist in the assessment of data quality tools; and
- Assist in the design of each data quality technique for improved data quality in a data warehouse.

Metadata Management

Table A-1: *Metadata Management – Data Quality Technique #1* addresses the themes, synonyms, and descriptions related to the data quality technique of metadata management. This table addresses the goals of the study in the following manner:

Clarification of terminology

As revealed in Table A-1, the key foundational metadata management themes of “data architecture” (Brackett, 1992, 1996, and 2000) and “metadata” (Brackett, 1992, 1996, and 2000; English, 1999; Kelly, 1995; Loshin, 2003; Redman, 2001) have many synonyms. Until uniform industry-standard terms emerge, this list of synonyms for data architecture and metadata may help to clarify terminologies. For IT data warehousing professionals interested in understanding the subtleties in definitions among synonyms such as "data architecture" (Brackett, 1992 and 1996) versus "common data architecture" (Brackett, 1992, 1996, and 2000) or "foredata" (Brackett, 1996) versus "afterdata" (Brackett, 1996), Table B-1: *Metadata Management Content Analysis* in Appendix B provides citations from the literature and page numbers for reference.

Profile of metadata management to assist in the assessment of tools

The presentation of categories of metadata should provide an IT data warehousing professional a good outline of metadata that must be managed. By listing the types of metadata contained in each category, the reader can better conceptualize the metadata comprising the category. The techniques to maintain metadata quality highlight areas where data quality tools can help manage metadata, in particular reverse engineering tools (Kimball & Caserta, 2004), data profiling tools (Eckerson, 2002; Kimball, et al., 1998; Kimball & Caserta, 2004; Loshin, 2003; Olson, 2003), data monitoring tools (Eckerson, 2002; Olson, 2003), and metadata management and quality tools (English, 1999; Huang, et al., 1999; Loshin, 2003).

Assistance in the design of metadata management for a data warehouse

Table A-1 contains an extensive list of metadata categories, types, and forms of metadata quality. The ability of an IT data warehousing team to manage all forms of corporate-wide metadata may be hindered by time and resource constraints. Nonetheless, the literature under review does indicate metadata management themes tailored specifically toward a data warehouse project. Notable highlights are:

- Loshin (2003) and Kimball, et al. (1998) suggest that the dimensional model is the most logical data relation diagram for a data warehouse.
- Front room metadata assists query tools and report writers accessing the data warehouse (Kimball, et al., 1998).
- Backroom metadata (Kimball, et al., 1998), reverse engineering (Kimball & Caserta, 2004), and data profiling (Olson, 2003) guide the design and maintenance of ETL processes.
- Kelly (1995) states that the data repository should contain metadata about the ETL process, particularly the source data to target data mappings.
- Analyzing the data warehouse data model for completeness and correctness is a form of metadata quality management (English, 1999).

Data Cleansing

Table A-2: *Data Cleansing – Data Quality Technique #2* addresses the themes, synonyms, and descriptions related to the data quality technique of data cleansing. This table addresses the goals of the study in the following manner:

Clarification of terminology

Table A-2 helps to clarify the definition of the key foundational theme of “data cleansing” by listing its many synonyms (English, 1999; Kimball & Caserta, 2004; Loshin, 2003; Redman, 1992 and 2001). While an industry-standard term is not found for the concept of data cleansing, most other data cleansing themes, fortunately, have limited synonyms and commonly used names.

Profile of data cleansing to assist in the assessment of tools

The concept of data cleansing is best understood by reviewing the data cleansing steps. When assessing data cleansing tools, an IT data warehousing professional can determine which data cleansing steps the tool under consideration supports. The literature under review highlights that most commercial data cleansing tools are focused on cleansing customer data (Eckerson, 2002; Olson, 2003), but Eckerson (2002) notes that emerging tool functionality will support non-customer data cleansing needs.

Assistance in the design of data cleansing for a data warehouse

Three key decisions emerge from the analysis of the literature regarding the use of data cleansing for a data warehouse. One consideration is whether data cleansing is most appropriate at the source system, during the ETL process, at the staging database, or within the data warehouse (Eckerson, 2002; Kimball & Caserta, 2004). Another decision is to decide what data elements to cleanse and how to cleanse them. If non-customer data is considered, the IT data warehousing professional must determine whether a tool exists on the market to cleanse the data. If not, custom software will be necessary. Finally, the IT data warehousing professional must be aware that not all data cleansing steps can be automated. The final two steps of analyzing data defect types (Eckerson, 2002; English, 1999) and preventing future errors (Eckerson, 2002; Kimball & Caserta, 2004; Redman, 1992, 1996, and 2001) require human intervention and analysis. As discovered in the information stewardship component, the role of data cleanup coordinator is an important role for data cleansing activities (Eckerson, 2002; English, 1999; Redman, 1992).

Information Stewardship

Table A-3: *Information Stewardship – Data Quality Technique #3* addresses the themes, synonyms, and descriptions related to the data quality technique of information stewardship. This table addresses the goals of the study in the following manner:

Clarification of terminology

Table A-3 highlights the many synonyms for terms in this component, most notably "data quality program" (Eckerson, 2002; Huang, et al., 1999; Olson, 2003; Redman, 1996 and 2001), "data quality council" (Eckerson, 2002; English, 1999; Olson, 2003; Redman, 2001), "detail data steward" (Brackett, 2000; Eckerson, 2002; English, 1999; Huang, et al., 1999; Kimball, et al., 1998; Kimball & Caserta, 2004; Redman, 1996), "data cleanup coordinator" (Eckerson, 2002; English, 1999; Kimball & Caserta, 2004; Redman, 1992), and "information quality analyst" (Eckerson, 2002; English, 1999; Kimball, et al., 1998; Kimball & Caserta, 2004). The disparity of terms may stem from the relative flexibility available to an organization to develop a data quality program. Eckerson (2002) outlines eight roles for the data quality team while English (1999) identifies six slightly different job functions. Another factor for the plethora of information stewardship synonyms is the authors' interchangeable use of "information" and "data" as prefixes to themes. For example, Redman (1996) uses the term "data quality program" (p. 18) and Huang, et al. (1999) refer to an "information quality program" (p. 27).

Profile of information stewardship to assist in the assessment of tools

The literature under review reveals that information stewardship is primarily a human resource endeavor. The role of tools is limited. Nonetheless, an IT data warehousing professional must consider how the data quality policy and guidelines will be communicated and distributed. An electronic means may be suitable. In addition, the data cleanup coordinator may utilize data cleansing tools outlined in Table A-2 (data cleansing), and the information quality analyst may use metadata quality reporting tools identified in Table A-1 (metadata management).

Assistance in the design of information stewardship for a data warehouse

Similar to metadata management, the IT data warehousing team may not possess the time and resources to institute a corporate-wide data quality program. Nevertheless, Olson (2003) points out that a data quality program is important "to create high-quality databases and maintain them at a high level" (p. 65). To achieve information stewardship objectives, the IT data warehousing professional should consider assigning the following roles:

- *Strategic data steward* to gain executive support (Eckerson, 2002; Redman, 1992, 1996, and 2001) and initiate (Olson, 2003) the data quality program.
- *Detail data steward* to maintain the data definition and resolve non-shared or redundant data (English, 1999).
- *Subject matter experts* to establish business accountability for information quality and strengthen the business and information systems partnership (English, 1999).

Appendices

Appendix A – Final Outcome of the Study

Three Data Quality Techniques: Themes, Synonyms & Definitions

Table A-1: Metadata Management - Data Quality Technique #1

Theme	Synonyms	Description
Data Architecture	Common data architecture; comprehensive data architecture; integrated data resource; data resource framework	The activities and framework related to identifying, naming, defining, structuring, maintaining quality, and documenting an enterprise data resource. A common data architecture is a formal and complete data architecture that provides context to the data resource so that it can be understood.
Metadata	Data definition; data documentation; data description; data resource data; data about data; foredata; afterdata	The definitions and documentation of the data architecture. Metadata describes and characterizes the data resource so that enterprise data can be easily understood, readily available and meaningful.
Categories of Metadata		
Meta-metadata		Data that describes the metadata, thereby providing a framework for developing new high-quality metadata.
Technical Metadata		Describes and characterizes the structure of data, how it is processed, and how it changes as it moves through

		the process.
Business Metadata		Describes data within a business context for greater value to the business customer.
Process Execution Metadata		Provides information about an ETL process, such as load time, rows loaded, and rows rejected.
Information Quality Measures		Provides quality statistics for data such as accuracy and timeliness so that business customers can gauge the quality of information derived from the data.
Types of Meta-metadata		
Data Naming Taxonomy		A common language for naming data that ensures unique names for all data in the common data architecture.
Data Naming Lexicon		Common words and abbreviations for the data naming taxonomy.
Data Thesaurus		Synonyms for data subjects and data characteristics.
Data Glossary		Definitions for words, terms, and abbreviations related to the common data architecture.
Data Translation Schemes		Translation algorithms and explanations for variations in data.
Data naming vocabulary		Common words with common meanings for data names.
Metadata Standards		Standards and practices for metadata, as established by a metadata organization or internal team.
Types of Technical Metadata		
Data Dictionary	Data product reference; data attribute structure	Formal names and definitions for data fields. The data dictionary may also include cross-references between the disparate data name and the common data name, definitions of changes that occur to the data field in upstream and downstream processes, and information about the accuracy of the data.
Data Structure	Enterprise model; common data structure; proper data structure; logical data structure; physical data structure	The proper logical and physical structure of data. The logical data structure represents how the data resource supports business activities. The physical data structure represents the structure of data in the manner that they are stored in files and databases. All data structures for an enterprise are referred to as the

		enterprise model.
Data Relation Diagram		The arrangement and relationships between data subjects. There are three types of data relation diagrams: subject relation diagram, file relation diagram, and entity-relation diagram.
Subject Relation Diagram		The arrangement and relationships between data subjects.
File Relation Diagram		The arrangement and relationships between data files in the physical data structure.
Entity-relation Diagram	Entity relationship diagram; entity-relationship modeling; entity-relationship model; entity relation diagram	The arrangement and relationships between data entities in the logical data structure.
Dimensional Model	Dimensional modeling	A form of logical data structure design more suitable for a data warehouse. The data entities are organized in a manner that is more intuitive than an Entity-relation Diagram and allows for high-performance data access.
Data Integrity	Data integrity rules	The formal definition of rules that ensure high-quality data in the data resource.
Data Profiling Metadata		A quantitative assessment of the values and quality of data in the data resource.
Types of Business Metadata		
Front Room Metadata		A more descriptive form of metadata that helps business customers to more easily query the data resource and write reports.
Data Clearinghouse	Data portfolio	Descriptions of data sources, unpublished documents, and projects related to the data resource. The data clearinghouse may also contain metadata about data that exist outside the organization. It is intended to support business activities.
Data Directory		Descriptions of organizations that maintain artifacts in the data clearinghouse and contacts in those organizations.
Business Rules	Data and data value rule enforcement; information compliance; business rules system	The rules that govern business processes. Ideally, knowledge about a business process is abstracted from the explicit implementation of the process and stored

		as metadata in a business rules system.
Types of Process Execution Metadata		
Back Room Metadata		ETL process metadata that guide the extraction, cleaning, and loading processes.
Types of Information Quality Measures		
Data Accuracy Measures		An objective measurement of a data sample against one or more business rules to determine its level of reliability and kind and degree of data errors.
Individual Assessment		A subjective measurement of how individuals within the organization perceive the quality of the information from the data resource.
Application Dependent Assessment		An objective measurement of how information quality may affect the organization.
Data Repository	Data resource guide; metadata repository; data resource library; metadata warehouse; metastore; metadata catalog; information library; repository; metadatabase	A database for metadata. The data repository provides an index to metadata for use by the organization. A data repository typically stores the data naming lexicon, data dictionary, data structure, data integrity, data thesaurus, data glossary, data product reference, data directory, data translation schemes, and data clearinghouse. A data repository for a data warehouse also holds details of the source-to-target mappings.
Metadata Quality		Metadata quality is critical for thorough understanding and utilization of the data resource.
Types of Metadata Quality		
Data Definition Quality		How well the data definition completely and accurately describes the meaning of enterprise data.
Data Standards Quality		How well the data standards enable people to easily define data correctly.
Data Name Quality		How well data is named in a way that clearly communicates its meaning.
Business Rule Quality		How well the business rules reflect the business policies.
Information and Data Architecture Quality		How well the information and data models are reused, stable, and flexible, and how well they meet the information needs of the organization.
Data Relationship Correctness		How well the relationships among entities and

		attributes in data models reflect the real-world objects and facts.
Business Information Model Clarity		How well the information model represents the business.
Operational Data Model Completeness and Correctness		How well the operational data model reflects the business processes.
Data Warehouse Data Model Completeness and Correctness		How well the data warehouse data model reflects the analytical needs of the business.
Techniques to Maintain Metadata Quality		
Data Resource Chain Management		Policies to ensure metadata standards are met for data sources obtained externally and in all levels of the internal data processing chain.
Data Refining	Information refining	Integration of undifferentiated raw data within the common data architecture into usable elemental units.
Reverse Engineering		Creation of an entity-relation diagram by reading the database data dictionary. Many data-profiling, data-modeling, and ETL tools offer this functionality.
Data Resource Survey and Inventory	Information needs analysis	The data resource survey and data resource inventory determine whether all of the information needed by the business customer is available in the data resource. A data resource survey gathers details from the business customer on their information needs. The data resource inventory is a detailed determination of the business customer's information needs and the data currently available in the data resource.
Data Profiling	Data auditing; data exploration prototype; data content analysis; anomaly detection phase; inferred metadata resolution	A process of analyzing data for the purpose of characterizing the information discovered in the data set. The purpose of data profiling is to identify data errors, create metrics to detect errors, and provide insight into how to resolve the data errors. Data profiling also validates the data definition by comparing existing data values to the intended data values. Data profiling can take the form of column property analysis to determine whether values in a data table are valid or invalid, structure analysis to determine structure rules and find structure violations, or data rules analysis to determine whether data in an

		entity meets the data rules intended for the entity.
Data Monitoring		Software programs that audit data at regular intervals or before data are loaded into another system like a data warehouse. The programs check for conformance to rules that may not be prudent to run at the transaction level or to verify whether data quality goals are being met.
Metadata Management and Quality Tools		Software tools that provide management of metadata, such as maintenance of business rules or data transformation rules. These tools utilize a rules engine that takes rules created with a rules language as input. Other types of metadata management tools assess conformance to metadata standards or provide a means to conduct a data resource survey.

Table A-2: Data Cleansing - Data Quality Technique #2

Theme	**Synonyms**	**Description**
Data Cleansing	Data cleaning; data cleanup; data correction; database cleanup; information product improvement; data reengineering; data scrubbing; data transformation; data-quality screen; cleaning and conforming	Data cleansing is the act of correcting missing or inaccurate data through error detection. Data cleansing entails elimination of duplicate records and filtering of bad data. Data cleansing can also entail the transformation of like data from disparate sources into a well-defined data structure (also known as conforming). For a data warehouse, data can be cleansed in the source database, during the Extract, Transform, & Load (ETL) process, in the staging area, or in the data warehouse directly.
Data Cleansing Steps		
Identify Data Sources		Determine which files or databases hold the data about an entity, which sources are most reliable, and which means is best to retrieve the data.
Extract and Analyze Source Data	Data auditing; source data validation; data profiling; data discovery	Extract representative data from the source files and discover characteristics and anomalies about the

		source data. The source data may also need to be "conditioned" to address obvious differences in the quality of the data.
Parse Data		Identify individual data elements within data fields and separate them into unique fields that have business meaning.
Standardize Data		Format the data based upon a specified standard or common library. This step may also entail expanding abbreviated fields to common, standard values.
Correct Data		Modify an existing incorrect value, modify a valid value to conform to a standard, or replace a missing value.
Enhance Data	Verification; data enrichment	Augment a record with additional attributes based upon an external library such as the United States Postal Service database for addresses.
Match and Consolidate Data	Deduplication; filtering; merge/purge; householding	Examine data to locate duplicate records for the same entity and then consolidate the data across the duplicates into a single "survivor" record. Consolidation may also include householding, which is the identification of customer records representing the same household.
Analyze Data Defect Types	Detect and report	Report patterns for defective data that were cleansed or not cleansed.
Prevent Future Errors		Prevention can take the form of education, process change, or new data edits. Data edits are automated routines that verify data values meet predetermined constraints upon entry into the system.
Data Cleansing Tools		Third-party software that assists with the examination, detection, and correction of data.
Data Cleansing Tool Functionality		Most data cleansing tools on the market are customer-centric. Tool strengths are data auditing, parsing, standardization, verification, matching, and consolidation/householding of name and address fields.
Emerging Data Cleansing Tool Functionality		
Parsing, standardization, and matching beyond name/address		Parsing, standardizing and matching algorithms applied to data fields other than name and address,

		such as email and product numbers.
Internationalization		International name and address data cleansing supporting extended character sets and international postal databases.
Data augmentation beyond the US postal database		Augmentation of geocoding, demographic, and psychographic data from information service provider databases.
Customer Key Managers		Use of internal match keys for matching of customers across time and systems.
Tool Integration		Ability to include data cleansing routines in other applications though an application interface (API).
Data integration Hubs	Real-time dimension manager system; real-time cleaning	A central repository that cleanses data real-time and publishes a standardized record.

Table A-3: Information Stewardship - Data Quality Technique #3

Theme	**Synonyms**	**Description**
Information Stewardship	Data resource management	Accountability for the quality of enterprise information through maintenance of the organization's data resource, using data engineering principles to ensure data quality.
Data engineering	Information engineering	The discipline for determining the true meaning of an organization's data and its information needs by designing, building, and maintaining a data resource library.
Information Stewardship Objectives		
Business accountability for information quality		Increase the value and quality of information and reduce poor information quality.
Maintenance of the data definition		Create and maintain a common definition of enterprise data to increase business communication, understanding, and productivity.

Resolution of non-shared or redundant data.		Conform disparate data to the common data definition by incorporating non-shared or redundant databases, interfaces, and applications into the shared data resource.
Strengthen the business and information systems partnership		Improve the effectiveness of information stewardship through a consensus-building, facilitated approach between the knowledgeable business and the information systems team members.
Data Quality Program	Data quality system; information quality program; data quality assurance program; data quality assurance initiative; data stewardship program	An initiative to implement data quality practices throughout an organization. A data quality program has clear objectives established in a data quality policy. Senior leadership must support the initiative. Business customer involvement in the data quality program helps ensure success. A successful data quality program has a management infrastructure and a data quality team. This team works to improve enterprise data and educate the organization on data quality.
Data Quality Policy	Data policy	A declaration of management responsibilities for data and information quality designed to outline the objectives of the data quality program and management accountabilities for achieving the objectives.
Information Stewardship Guidelines		A document defining the roles and responsibilities of the information steward, and guidelines for implementing data quality processes. Along with the data quality policy and training, the information stewardship guidelines are the support tools for information stewards.
Executive Support for a Data Quality Program		A data quality program must have the support of senior management, ideally initiated by the CEO, to ensure long-term success. A data quality program should be managed by a chief data quality officer or by executives in each business area.
Business Customer Involvement in the Data Quality Program	Data definition team; business-oriented information engineering	Knowledgeable business experts must be involved in the data quality program. Facilitated sessions with representatives from all business areas are critical to

		develop the common metadata. Business customers are also ultimately directly involved in the implementation of business rules and processes to improve data quality.
Information Stewardship Team		The information stewardship team is comprised of two bodies: the data quality council and the data quality team.
Data Quality Council	Executive information steering team; corporate stewardship committee; data quality assurance advisory group	The data quality council is a senior management body that ensures the data quality policy is carried out. It oversees the activities of the data quality team and gives authority to the team members to carry out their responsibilities.
Data Quality Team	Data quality assurance department; business information stewardship team	The data quality team executes the data quality responsibilities described in the data quality policy and the information stewardship guidelines. The team is typically comprised of these roles (or combinations thereof): a chief quality officer, strategic data stewards, tactical data stewards, detail data stewards, data cleanup coordinators, information quality analysts, information quality process improvement facilitators, information quality training coordinators, and subject matter experts.
Chief Quality Officer		The senior officer who oversees the data quality program.
Strategic Data Steward	Data quality leader; information quality manager	The executive who manages the data quality team. This person has decision-making authority for implementing the data quality program, building organizational awareness, and committing resources.
Tactical Data Steward		In very large organizations, the tactical data steward acts as a liaison between the strategic data steward and the detail data stewards spread across global sites.
Detail Data Steward	Information architecture quality analyst; information steward; data steward; data guardian; data custodian; data coordinator; data analyst; data trustee; data curator; data administrator; data facilitator; data negotiator; data interventionist; information product	A person knowledgeable about the data resource. The detail data steward is responsible for the data definition, data model, metadata, and overall data quality. This person coordinates information processes to ensure delivery of quality information to the business consumer. The detail data steward is also

	manager	responsible for establishing information and data quality metrics that will improve data quality.
Data Cleanup Coordinator	Information-quality leader; data quality tools specialist; data keeper	The data cleanup coordinator is responsible for data cleansing tasks. This person also performs operational activities to detect and resolve data quality issues. The data cleanup coordinator may be responsible for the data dictionary.
Information Quality Analyst	Data warehouse quality assurance analyst; data-quality specialist; data quality analyst	The information quality analyst is responsible for auditing, monitoring, and measuring data quality in an operational capacity. This person reports on the results of the measurements and resolves data quality issues.
Information Quality Process Improvement Facilitator	Process improvement facilitator	This person facilitates efforts to reengineer business processes for resolution of ongoing data quality issues.
Subject Matter Expert		The subject matter expert is typically a knowledgeable business analyst whose understanding of the business is necessary to understand data, define business rules, and measure data quality.

Appendix B – Results of Content Analysis

Table B-1: Metadata Management Content Analysis

Reference Key	
BR00	Brackett (2000)
BR94	Brackett (1994)
BR96	Brackett (1996)
EC02	Eckerson (2002)
EN99	English (1999)
HU99	Huang, et al. (1999)
KE95	Kelly (1995)
KI04	Kimball & Caserta (2004)
KI98	Kimball, et al. (1998)
LO03	Loshin (2003)
OL03	Olson (2003)
RE01	Redman (2001)
RE92	Redman (1992)
RE96	Redman (1996)

Description	*Reference*	*Page Number*	*Theme/Synonym*
Data Architecture Defined			
data architecture - "contains all the activities related to describing, structuring, maintaining quality, and documenting the data resource…The Data Architecture component contains four activities: Data Description…Data Structure…Data Quality…Data Documentation"	BR92	28	Theme
data architecture - "The component of the data resource framework that contains all activities, and the products of those activities, related to the identification, naming, definition, structuring, quality, and documentation of the data resource for an organization."	BR96	56	Synonym
common data architecture - "is a data architecture that provides a common context within which all data are defined to determine their true content and meaning so they can be integrated into a formal data resource and readily shared to support information needs. It is consistent across all data so they can be refined within a common context. It is a common base for formal naming, comprehensive definition, proper structuring, maintenance of quality, and complete documentation of all data."	BR92	31	Synonym
common data architecture - "is a formal, comprehensive data architecture that provides a common context within which all data are understood and integrated."	BR00	15	Synonym
common data architecture - "is a formal, comprehensive data architecture that provides a common context within which an integrated data resource is developed so that it adequately supports the business information demand."	BR96	57	Synonym

comprehensive data architecture - "the concept of a total infrastructure for information technology, the establishment of a data resource framework within that infrastructure, and the definition of a formal data architecture within that framework."	BR96	51	Synonym
integrated data resource - "is a data resource where all data are integrated within a common context and are appropriately deployed for maximum use in supporting the business information demand…High-quality metadata adequately describe the data resource and are readily available to clients so they can easily identify and readily access any data needed to perform their business activities."	BR96	36	Synonym
data resource framework - "represents a discipline for the complete development and maintenance of an integrated data resource. It three components are data management, data architecture, and data availability…"	BR96	55	Synonym

Metadata Defined

metadata - "a catalog of the intellectual capital that surrounds the creation, management, and use of a collection of information."	LO03	84	Theme
metadata - "is the term which is used to describe the definitions of the data that is stored in the data warehouse."	KE95	141	Synonym
data definition - "refers to the set of information that describes and defines the meaning of the 'things' and events, called entity types, the enterprise should know about and what facts, called attributes, it should know about them to accomplish its mission. The term data definition as used here refers to all of the descriptive information about the name, meaning, valid values, and business rules that govern its integrity and correctness, as well as the characteristics of data design that govern the physical databases...The term data definition as used here is synonymous with the technical term metadata, which means 'data that describes and characterizes other data'."	EN99	84-85	Synonym
data definition - "is a formal data definition that provides a complete, meaningful, easily read, readily understood, real-world definition of the true content and meaning of data. Comprehensive data definitions are based on sound principles and a set of guidelines. These ensure that they provide enough information to clients so the formal data resource can be thoroughly understood and fully utilized to meet information needs."	BR92	68	Synonym
metadata - "are the data describing the foredata. They are the afterdata that provide definitions about the foredata, including the foredata that describe objects and events and data about the quality of data describing objects and events."	BR96	190	Synonym
robust data documentation - "is documentation about the data resource that is complete, current, understandable, non-redundant, readily available, and known to exist. Achieving robust data documentation requires a new approach to designing and managing data documentation...Documentation about the data resource is often referred to as metadata, which is commonly defined as data about the data."	BR00	149	Synonym
comprehensive data definition - "is a formal data definition that provides a complete, meaningful, easily read, readily understood definition that thoroughly explains the content and meaning of the data. It helps people thoroughly understand the data and use the data resource efficiently and effectively to meet the business information demand."	BR00	63	Synonym
data description - "ensures the formal naming and comprehensive definition of all data."	BR92	28	Synonym
data description - "includes the formal naming and comprehensive definition of data."	BR96	69	Synonym
data documentation - "ensures current, complete, continuing documentation of the entire data architecture component."	BR92	28	Synonym
data resource data - "are data that describe the data resource…They are more commonly called 'metadata' (data about data)."	RE01	28	Synonym
foredata - "are the upfront data that describe those objects and events. Foredata are the data that people use to track or manage objects and events in the real world. Foredata include both data representing the objects and events and data about the quality of the data representing the objects and events."	BR96	190	Synonym
common metadata - "are metadata developed within the common data architecture to provide all the detail necessary to thoroughly understand the data resource and how it can be improved to meet the business information demand."	BR96	192	Theme

metadata demand - "People are documenting data in CASE tools, data dictionaries, repositories, text processors, spreadsheets, and a variety of other products. It is difficult to find all the metadata and to integrate those metadata for a consistent understanding of the real data...The metadata demand is an organization's need for complete, accurate data about its data resource that is easily understandable and readily available to anyone using, or planning to use, that data resource."	BR96	13-14	Theme

Categories of Metadata

meta-metadata - "are the data describing the metadata. They are the data that provide the framework for developing high-quality metadata."	BR96	190	Theme
categories of metadata for ETL - "Business metadata…Technical metadata…Process execution metadata"	KI04	357	Theme
two areas of metadata - "technical metadata, which describes the data mechanics, and business metadata, which describes the business perception of that same information."	LO03	85	Synonym
technical metadata - "describes the structure of information, whether it is the data that is sourcing the warehouse or the data in the warehouse. Technical metadata characterizes the structure of data, the way that data move, and how it is transformed as it moves from one location to another."	LO03	85	Theme
technical metadata - "Representing the technical aspects of data, including attributes such as data types, lengths, lineage, results from data profiling, and so on"	KI04	357	Synonym
business metadata - "Describing the meaning of data in a business sense"	KI04	357	Theme
business metadata - "incorporates much of the same information as technical metadata, as well as: *Metadata that describes the structure of data as perceived by business clients; * Descriptions of the methods for accessing data for client analytical applications; * Business meanings for tables and their attributes; * Data ownership characteristics and responsibilities; * Data domains and mappings between those domains, for validation; * Aggregation and summarization directives; * Reporting directives; * Security and access policies; * Business rules"	LO03	88	Synonym
process execution metadata - "Presenting statistics on the results of running the ETL process itself, including measures such as rows loaded successfully, rows rejected, amount of time to load, and so on"	KI04	357	Theme
information quality measures - "Information quality characteristics, such as accuracy and timeliness, are the aspects or dimensions of information quality important to knowledge workers…Information quality measures are the information quality characteristics assessed."	EN99	141	Theme

Types of Meta-metadata

data naming taxonomy - "provides unique names for all logical and physical data within the common data architecture."	BR92	52	Theme
data naming taxonomy - "provides a common language for naming data."	BR96	72	Synonym
formal data naming taxonomy - "was developed to provide a primary name for all existing and new data, and all components in the data resource. The data naming taxonomy also provides a way to uniquely designate other features in the data resource, such as data characteristic substitutions and data values."	BR00	37	Synonym
data naming lexicon - "contains common words and word abbreviations for the data naming taxonomy in the common data architecture."	BR96	196	Theme
data thesaurus - "contains synonyms for data subjects and data characteristics in the common data architecture."	BR96	196	Theme
data glossary - "contains definitions for words, terms, and abbreviations related to the common data architecture."	BR96	196	Theme
data translation schemes - "are translation algorithms and explanations for data variations in the common data architecture."	BR96	196	Theme
data naming vocabulary - "provides common words with common meanings for all data names."	BR92	54	Theme
metadata standards - "Many organizations attempt to standardize metadata at various levels…To maintain manageable jobs for all of your enterprise data warehouse ETL processes, your data warehouse team must establish standards and practices for the ETL team to follow."	KI04	377-378	Theme

Types of Technical Metadata

Data Dictionary

data dictionary - "includes formal names and comprehensive definitions for all data in the common data architecture."	BR96	196	Theme
comprehensive data dictionary - "should provide definitions of stored data fields. In addition, it should provide definitions of all data fields in processes upstream of the database and the changes to these fields downstream."	RE92	244	Synonym
data product reference - "is inventory of existing data, including definitions, structure, integrity, and cross references to the common data architecture."	BR96	196	Synonym
data attribute structure - "is a list that shows the data attributes contained within a data entity and the roles played by those data attributes."	BR00	93	Synonym
formal data name - "readily and uniquely identifies a fact or group of facts in the data resource. It is developed within a formal data naming taxonomy and is abbreviated, when necessary, with a formal set of abbreviations and an abbreviation algorithm."	BR00	36-37	Theme
data cross-reference - "is a link between disparate data names and common data names."	BR96	239	Theme
data cross-reference - "Cross referencing disparate data to the common data architecture is a major step in understanding and managing disparate data."	BR92	257	Synonym
"Data accuracy is documented in both the data name and the data description."	BR92	147	Theme

Data Structure

data structure - "is the structure for all data in the common data architecture."	BR96	196	Theme
data structure - "ensures the proper logical and physical structure of data."	BR92	28	Synonym
common data structure - "is the structure of data within the common data model that provides a full understanding of all the disparate data structures and multiple perspectives of the real world those data structures represent."	BR96	102-103	Synonym
proper data structure - "is a data structure that provides a suitable representation of the business, and the data resource supporting that business, that is relevant to the intended audience…A proper data structure consists of an entity-relation diagram and an attribute structure."	BR00	91-92	Synonym
enterprise model - "will comprise a number of separate models which, combined together, provide an integrated picture of the enterprise. There may be many of these separate models which describe the enterprise in terms of enterprise strategy, enterprise organization, enterprise data, enterprise processes, or enterprise culture."	KE95	61	Synonym
logical data structure - "is a data structure representing logical data. It is generally developed to show how the formal data resource supports business activities."	BR92	92	Theme
logical data structure - "is the structure of data in the logical data model."	BR96	103	Synonym
physical data structure - "is a data structure representing physical data. It is generally developed from a logical data structure to show how data are physically stored in files and databases."	BR92	92	Theme
physical data structure - "is the structure of data in the physical data model."	BR96	103	Synonym

Data Relation Diagram

data relation diagram - "shows the arrangement and relationship of data subjects in the common data architecture, but does not show any contents of a data subject."	BR92	92	Theme
data relation diagram - "refers to a set of three diagrams representing the three types of data models." (subject relation diagram, file relation diagram, entity-relation diagram)	BR96	109	Synonym
subject relation diagram - "shows the arrangement and relationship of data subjects in the common data structure."	BR96	113	Theme
file relation diagram - "represents the arrangement and relationship of data files for the physical data model."	BR96	114	Theme
entity-relation diagram - "contains only the data entities and the data relations between those data entities."	BR00	92	Theme
entity relationship diagram - "To identify the (high-level) entities which occur in an enterprise and to define the relationships which exist between the entities."	KE95	75	Synonym
entity-relationship modeling - "is a logical design technique that seeks to eliminate data redundancy."	KI98	140	Synonym

entity-relationship model - "a reasonable scheme for mapping a business process to a grouped sequence of table operations to be executed as a single unit of work."	LO03	77	Synonym
entity relation diagram - "represents the arrangement and relationship of data entities for the logical data structure."	BR96	110	Synonym
dimensional model - "an alternate technique to model data has evolved that allows for information to be represented in a way that is more suitable to high-performance access….is a much more efficient representation for data in a data warehouse."	LO03	79	Theme
dimensional modeling - "is a logical design technique that seeks to present the data in a standard framework that is intuitive and allows for high-performance access."	KI98	144	Synonym

Data Integrity

data integrity - "contains rules for all data in the common data architecture."	BR96	196	Theme
data integrity - "is the formal definition of comprehensive rules and the consistent application of those rules to ensure high quality data in the formal data resource. It deals with how well data are maintained in the formal data resource. It is both an indication of how well data are maintained in the formal data resource and an activity to ensure that the formal data resource contains high-quality data."	BR92	129	Synonym
data integrity - "is the formal definition of comprehensive rules and the consistent application of those rules to ensure high integrity data."	BR96	145	Synonym
data integrity - "Dr. Edgar F. Codd proposed five integrity rules that must be followed by any true relational database management system…Simply put, Codd's integrity rules ensure data meet specifications demanded by the designer and the user."	HU99	63	Synonym
precise data integrity rule - "is a data integrity rule that precisely specifies the criteria for high-quality data values and reduces or eliminates data errors. The consistent application and enforcement of those rules ensure high-quality data values."	BR00	121	Theme

Data Profiling Metadata

data profiling metadata - "Good data-profiling analysis takes the form of a specific metadata repository describing…a good quantitative assessment of your original data sources."	KI04	125	Theme

Types of Business Metadata

Front Room Metadata

front room metadata - "The front room metadata is more descriptive, and it helps query tools and report writers function smoothly."	KI98	435	Theme

Data Clearinghouse

data clearinghouse - "contains descriptions of data sources, unpublished documents, and projects related to the data resource."	BR96	196	Theme
data portfolio - "is meta-data about the data that exist inside and outside the organization that can be accessed and used to support business activities. A comprehensive data portfolio is developed through general data surveys and detailed data inventories."	BR92	332	Synonym

Data Directory

data directory - "contains descriptions of organizations maintaining data sources, unpublished documents, and data projects and contacts in those organizations."	BR96	196	Theme

Business Rules

business rules - "business processes are governed by a set of business rules."	LO03	92	Theme
data and data value rule enforcement - "Data and value rules range from simple business rules…to more complex logical checks."	KI04	135	Synonym
information compliance - "is a concept that incorporates the definition of business rules for measuring the level of conformance of sets of data with client expectations. Properly articulating data consumer expectations as business rules lays the groundwork for both assessment and ongoing monitoring of levels of data quality."	LO03	140	Synonym

business rules system - "all knowledge about a business process is abstracted and is separated from the explicit implementation of that process."	LO03	92	Synonym

Types of Process Execution Metadata

Back Room Metadata

back room metadata - "The back room metadata is process related, and it guides the extraction, cleaning, and loading processes."	KI98	435	Theme

Types of Information Quality Measures

IQ metrics - "the IPM must have three classes of metrics: * Metrics that measure an individual's subjective assessment of IQ (how good do people in our company think the quality of our information is) * Metrics that measure IQ quality along quantifiable, objective variables that are application independent (how complete, consistent, correct, and up to date the information in our customer information system is) * Metrics that measure IQ quality along quantifiable objective variables that are application dependent (how many clients have exposure to the Asian financial crisis that our risk management system cannot estimate because of poor quality information). Used in combination, metrics from each of these classes provides fundamental information that goes beyond the static IQ assessment to the dynamic and continuous evaluation and improvement of information quality.	HU99	60-61	Theme
metrics - "One use is to demonstrate to management that the process is finding facts…Metrics can be useful to show improvements…Another use of metrics is to qualify data…Metrics can then be applied to generate a qualifying grade for the data source…The downside of metrics is that they are not exact and they do not solve problems."	OL03	83	Theme
data accuracy measurements - "Organizations just starting out do not need sophisticated, scientifically defensible measurements. They need simple measures that indicate where they are, the impact(s), and the first couple of opportunities for improvement...data accuracy measurements are essential and a good place to start<.>"	RE01	108,110	Theme
forms of information quality measurements - "Data assessment is composed of two forms of quality inspection. The first form of assessment is automated information quality assessment that analyzes data for conformance to the defined business rules. The second is a physical information quality assessment to assure the accuracy of data by comparing the data values to the real-world objects or events the data represents. <The data assessment> objective is to measure a data sample against one or more quality characteristics in order to determine its level of reliability and to discover the kind and degree of data defects."	EN99	177	Theme

Data Repository

data repositories - "are specially designed databases for data resource data."	RE01	173	Theme
data resource guide - "A comprehensive data resource guide provides extensive information about all data in the data resource library. It is an information system that maintains meta-data about the formal data resource."	BR92	17	Synonym
metadata repository - "The primary software tool for managing data quality is the metadata repository."	OL03	19	Synonym
data resource library - "is a library of data for an organization…"	BR96	44	Synonym
metadata warehouse - "provides an index to the data in the data resource library just like a card catalog provides an index to the works in a library."	BR96	44	Synonym
metadata warehouse - "goes beyond traditional data dictionaries, data catalogues, and data repositories to provide a personal help desk for increasing the awareness and understanding of the data resource. It provides a usable, understandable index to the data resource supported by client-friendly search routines."	BR96	193	Synonym
metastore - "holds the metadata, needs to identify the 'pedigree' of the data in the data warehouse I.e. the quality, origin, age, and integrity of the data…It is also important for the metastore to hold details of the transformation process, (where data is mapped from the source systems to the data warehouse), so that the users can reverse engineer the derived and summary data into the original components."	KE95	142	Synonym
metadata catalog - "Terms like information library, repository, and metadatabase, among others, have all been used to describe this data store…In the best of all possible worlds, the metadata catalog would be the single, common storage point for information that drives the entire warehouse process."	KI98	445	Synonym

metadata warehouse components -"data naming lexicon…data dictionary…data structure…data integrity…data thesaurus…data glossary…data product reference…data directory…data translation schemes…data clearinghouse"	BR96	196-197	Theme

Metadata Quality

quality metadata - "are critical for thoroughly understanding and fully utilizing an integrated data resource."	BR96	185	Theme

Types of Metadata Quality

data definition quality - "How well data definition completely and accurately describes the meaning of the data the enterprise needs to know."	EN99	88	Theme
data standards quality - "The data standards enable people to easily define data completely, consistently, accurately, clearly, and understandably."	EN99	87	Synonym
business rule quality - "How well the business rules specify the policies that govern business behavior and constraints."	EN99	88	Theme
information and data architecture quality - "How well information and data models are reused, stable, and flexible and how well they depict the information requirements of the enterprise; and how well the databases implement those requirements and enable capture, maintenance, and dissemination of the data among the knowledge workers."	EN99	88	Theme
data name quality - "Data is named in a way that clearly communicates the meaning of the objects named."	EN99	87	Theme
data relationship correctness - "The specification of relationships among entities and attributes in data models accurately reflects the correct nature of relationships among the real-world objects and facts."	EN99	88	Theme
business information model clarity - "The high-level information model represents and communicates the fundamental business resources or subjects, and fundamental business entity types the enterprise just know about completely and clearly."	EN99	88	Theme
data model completeness and correctness for operational data - "The data model of operational data reflects completely all fact types required to be known by the enterprise to support all business processes and all business or functional areas. This detailed model correctly illustrates the relationships among entity types and between entity types and their descriptive attributes."	EN99	89	Theme
data warehouse model completeness and correctness to support strategic and decision processes - "The data model of strategic or tactical information (for data warehouses or data marts) completely and accurately reflects the information requirements to support key decisions, trend analysis, and risk analysis required to support the planning and strategic management of the enterprise."	EN99	89	Theme

Techniques to Maintain Metadata Quality

Data Resource Chain Management

data resource chain - "In most organizations, the data resource data are not up to the standards suggested by the library. For data obtained from the outside, supplier management should extend to data resource data as well. An internally, apply information chain management to implement a high-level resource data chain...Implement an end-to-end data resource chain to ensure that data resource data are well-defined, kept up-to-date, and made easily available to all. Implement data modeling and standards chains as support."	RE01	30,33	Synonym

Data Refining

data refining - "integrates disparate data within the common data architecture to support the business information demand."	BR96	224	Theme
information refining - "is a process that takes undifferentiated raw data, extracts the content into elemental units, and recombines those elemental units into usable information."	BR92	14	Synonym

Reverse Engineering

reverse engineering - "is a technique where you develop an ER diagram by reading the existing database metadata. Data-profiling tools are available to make this quite easy. Just about all of the standard data-modeling tools provide this feature, as do some of the major ETL tools."	KI04	67	Theme

Data Resource Survey and Inventory

data completeness - "ensures that all data necessary to meet the business information demand are available in the data resource. Data completeness is managed through data resource surveys and data resource inventories."	BR96	175	Theme
data resource survey - "consists of a data availability survey, a data needs survey, and a data survey analysis. It provides information about broad groupings of data needed to support an organization's business strategies and broad groupings of data that currently exist."	BR92	334	Theme
data resource survey - "is a high-level determination of an organization's data needs the data available to the organization based on a higher level data classification scheme."	BR96	175	Synonym
information needs analysis - "To provide a guide at a strategic level and at an operational level what are the key information needs of the key decision makers."	KE95	78	Synonym
data resource inventory - "consists of a data availability inventory, a data needs inventory, and a data inventory analysis. It provides detailed information about what data are needed to support business activities and what data currently exist."	BR92	338	Theme
data resource inventory - "is a detailed determination of the organization's data needs and the data available to the organization based on data subjects and data characteristics."	BR96	176	Synonym

Data Profiling

data profiling - "has emerged as a major new technology. It employs analytical methods for looking at data for the purpose of developing a thorough understanding of the content, structure, and quality of the data…Data profiling uses two different approaches for assessing data quality. One is discovery, whereby processes examine the data and discover characteristics from the data without the prompting of the analyst...The second approach is assertive testing. The analyst poses conditions he believes to be true about the data and then executes data rules against the data that check for these conditions to see if it conforms or not."	OL03	20	Theme
data profiling - "to discover metadata when it is not available and to validate metadata when it is available. Data profiling is a process of analyzing raw data for the purpose of characterizing the information embedded within a data set."	LO03	109	Synonym
data auditing - "<aka> profiling. The purpose of the assessment is to (1) identify common data defects (2) create metrics to detect defects as they enter the data warehouse or other systems, and (3) create rules or recommend actions for fixing the data."	EC02	19	Synonym
data exploration prototype - "To better understand actual data content, a study can be performed against current source system data."	KI98	303	Synonym
data content analysis - "Understanding the content of the data is crucial for determining the best approach for retrieval. Usually, it's not until you start working with the data that you come to realize the anomalies that exist within it."	KI04	71	Synonym
anomaly detection phase - "A data anomaly is a piece of data that does not fit into the domain of the rest of the data it is stored with."	KI04	131	Synonym
inferred metadata resolution - "discovering what the data items really look like and providing a characterization of that data for the next steps of integration."	LO03	110	Synonym
data profiling inputs - "There are two inputs: metadata and data. The metadata defines what constitutes accurate data…However, the metadata is almost always inaccurate and incomplete. This places a higher burden on attempts to use it with the data. Data profiling depends heavily on the data. The data will tell you an enormous amount of information about your data if you analyze it enough."	OL03	124	Theme
data profiling outputs - "The primary output of the data profiling process is best described as accurate, enriched metadata and facts surrounding discrepancies between the data and the accurate metadata. These facts are the evidence of inaccurate data and become the basis for issues formation and investigation."	OL03	129	Theme
data profiling for column property analysis - "Analysis of column properties is the process of looking at individual, atomic values and determining whether they are valid or invalid. To do this, you need a definition of what is valid. This is in the metadata. It consists of a set of definitional rules to which the values need to conform."	OL03	143	Theme

data profiling for structure analysis - "There are two issues to look for in structure analysis. One is to find violations to the rules that should apply. This point to inaccurate data. The other is to determine and document the structure rules of the metadata. This can be extremely valuable when moving data, mapping it to other structures, or merging it with other data."	OL03	173	Theme
data profiling for data rules - "Data rules are specific statements that define conditions that should be true all of the time. A data rule can involve a single column, multiple columns in the same table, or columns that cross over multiple values. Rules can also be restricted to the data of a single business object or involve data that encompasses sets of business objects."	OL03	215	Theme
data profiling tools strengths - "is intended to complete or correct the metadata about source systems. It is also used to map systems together correctly. The information developed in profiling becomes the specification information that is needed by ETL and data cleansing products."	OL03	53	Theme

Data Monitoring

data monitoring - "A data monitoring tool can be either transaction oriented or database oriented. If transaction oriented, the tool looks at individual transactions before they cause database changes. A database orientation looks at an entire database periodically to find issues."	OL03	20	Theme
monitor data quality - "companies need to build a program that audits data at regular intervals, or just before or after data is loaded into another system such as a data warehouse. Companies then use audit reports to measure their progress in achieving data quality goals and complying with service level agreements negotiated with business groups.	EC02	24	Synonym
data monitoring benefits - "the addition of programs that run periodically over the databases to check for the conformance to rules that are not practical to execute at the transaction level. They can be used to off-load work from transaction checks when the performance of transactions is adversely affected by too much checking. Because you can check for more rules, they can be helpful in spotting new problems in the data that did not occur before."	OL03	96	Theme

Metadata Management and Quality Tools

metadata management and quality tools - "Management and control tools that provide quality management of metadata, such as definition and control of business rules, data transformation rules, or provide for quality assessment or control of metadata itself, such as conformance to data naming standards."	EN99	313-314	Theme
information quality analysis tools - "Analysis tools that extract data from a database or process, measure its quality, such as validity or conformance to business rules, and report its analysis."	EN99	312	Synonym
business rule discovery tools - "Rule discovery tools that analyze data to discover patterns and relationships in the data itself. The purpose is to identify business rules as actually practiced by analyzing patterns in the data."	EN99	312	Theme
IQ survey tool - "To perform the necessary IQ analysis efficiently and effectively, however, it would be useful to have some computer-based tools to facilitate the analysis.	HU99	66	Theme
rules language - "All rules-based systems employ some kind of rules language as a descriptive formalism for describing all the aspects of the business process, including the system states, the actors, the inputs and events, the triggers, and the transitions between states."	LO03	102	Theme
rules engine - "is an application that takes as input a set of rules, creates a framework for executing those rules, and acts as a monitor to a system that must behave in conjunction with those rules."	LO03	103	Theme

Table B-2: Data Cleansing Content Analysis

Reference Key	
BR00	Brackett (2000)
BR94	Brackett (1994)
BR96	Brackett (1996)
EC02	Eckerson (2002)
EN99	English (1999)
HU99	Huang, et al. (1999)
KE95	Kelly (1995)
KI04	Kimball & Caserta (2004)
KI98	Kimball, et al. (1998)
LO03	Loshin (2003)
OL03	Olson (2003)
RE01	Redman (2001)
RE92	Redman (1992)
RE96	Redman (1996)

Description	*Reference*	*Page Number*	*Theme/Synonym*
Data Cleansing Defined			
data cleansing - "The terms cleansing, cleaning, cleanup, and correcting data are used synonymously to mean correcting missing and inaccurate data values."	EN99	237	Theme
database cleanups - "are distinguished from everyday editing in that cleanups are usually conducted outside the scope of everyday operations. Most database cleanups are simply sophisticated error detection, error localization, and error correction routines."	RE92	249	Synonym
database clean-ups - "There are any number of good computer tools that can automate error detection and many Information Technology departments are skilled at using them. Error correction is more problematic, but it can often be farmed out to relatively low-paid temps...Finally, while data clean-up is certainly not easy, the job can be fairly well delineated and completed in a reasonable amount of time."	RE01	54-55	Synonym
data cleansing - "A large part of the cleansing process involves identification and elimination of duplicate records; much of this process is simple, because exact duplicates are easy to find…The difficult part of eliminating duplicates is finding those nonexact duplicates - for example, pairs of records where there are subtle differences in the matching key."	L03	135	Synonym
data cleansing synonyms - "Information product improvement, basically the correction of defective data, is sometimes called data reengineering, data cleansing, data scrubbing, or data transformation."	EN99	237	Synonym
data reengineering - "implies the transformation of unarchitected data into architected and well-defined data structures."	EN99	237	Synonym
data-quality screen - "is physically viewed by the ETL team as a status report on data quality, but it's also a kind of gate that doesn't let bad data through."	KI04	114	Synonym

cleaning and conforming - "actually changes data and provides guidance whether data can be used for its intended purposes."	KI04	113	Synonym
conforming - "Integration of data means creating conformed dimension and fact instances built by combining the best information from several data sources into a more comprehensive view. To do this, incoming data somehow needs to be made structurally identical, filtered of invalid records, standardized in terms of its content, deduplicated, and then distilled into the new conformed image."	KI04	148	Theme
options for cleaning data - "* Cleanse data at the source. * Transform data in the ETL."	KI04	406	Theme
places to clean data - "at the source…in a staging area…ETL process…in the data warehouse"	EC02	24	Synonym
Data Cleansing Steps			
reengineer and cleanse data process steps - "Identify Data Sources…Extract & Analyze Source Data…Standardize Data…Correct and Complete Data…Match and Consolidate Data…Analyze Data Defect Types	EN99	245-246	Theme
data cleansing phases - "Parsing…Standardization…Abbreviation Expansion…Correction…Updating Missing Fields"	L03	136-139	Theme
data cleansing methods - "Correct…Filter…Detect and Report…Prevent"	EC02	22-23	Theme
Identify Data Sources			
identify data sources - "documents all pertinent files from all files that may hold data about a given entity, and determines which is most authoritative, if any, and where to cleanse or extract data for conversion or propagation to a target database or data warehouse."	EN99	247	Theme
Extract and Analyze Source Data			
extract and analyze source data - "extracts representative data from the source files and analyzes it to confirm that the actual data is consistent with its definition and to discover any anomalies in how the data is used and what it means. This uncovers new entity types, attributes, and relationships that may need to be included in the target data architecture."	EN99	250	Theme
data auditing - "Also called data profiling or data discovery, these tools or modules automate source data analysis. They generate statistics about the content of data fields."	EC02	27	Synonym
validating the source data - "other items of data on operational systems <that> are not intrinsic to the operational process and may have fallen into some decay…may have to be tackled before proceeding to migrate the data."	KE95	135	Synonym
conditioning the source data - "Because there will be considerable differences in the quality of the data on different operational systems it will be necessary in some instances to 'condition' the data on the operational systems before it is transported in the data warehouse environment."	KE95	134	Theme
Parse Data			
parsing - "is the process of identifying meaningful tokens within a data instance and then analyzing token streams for recognizable patterns. A token is a conglomeration of a number of single words that have some business meaning."	L03	136	Theme
parsing - "Parsing locates and identifies individual data elements in customer files and separates them into unique fields."	EC02	27	Synonym
Standardize Data			
standardize data - "This process standardizes data into a sharable, enterprise wide set of entity types or attributes."	EN99	252	Theme
standardization - "is the process of transforming data into a form specified as a standard."	L03	136	Theme
standardization - "Once files have been parsed, the elements are standardized to a common format defined by the customer…Standardization makes it easier to match records. To facilitate standardization, vendors provide extensive reference libraries, which customers can tailor to their needs. Common libraries include lists of names, nicknames, cardinal and ordinal numbers, cities, states, abbreviations, and spellings."	EC02	27	Synonym
abbreviation expansion - "Abbreviations must be parsed and recognized, and then a set of transformational business rules can be used to change abbreviations into their expanded form."	L03	137	Theme

Correct Data

Correct - "Most cleansing operations involve fixing both defective data elements and records. Correcting data elements typically requires you to (1) modify an existing incorrect value (e.g. fix a misspelling or transposition), (2) modify a correct value to make it conform to a corporate or industry standard (e.g. substitute 'Mr.' for 'Mister', or (3) replace a missing value. You can replace missing values by either inserting a default value (e.g. "unknown") or a correct value from another database, or by asking someone who knows the correct value. Correcting records typically requires you to (1) match and merge duplicate records that exist in the same file or multiple files, and (2) decouple incorrectly merged records. Decoupling is required when a single record contains data describing two or more entities, such as individuals, products, or companies.	EC02	22	Theme
correct and complete data - "improves the quality of the existing data by correcting inaccurate or nonstandardized data values, and finding and capturing missing data values."	EN99	257	Synonym
correction - "Once components of a string have been identified and standardized, the next stage of the process attempts to correct those data values that are not recognized and to augment correctable records with the corrected information."	L03	138	Synonym
updating missing fields - "one aspect of data cleansing is being to fill fields that are missing information…Given the corrected data, the proper value may be filled in. For unknown attributes, the process of cleansing and consolidation may provide the missing value."	L03	139	Theme

Enhance Data

data enhancement - "is a process to add value to information by accumulating additional information about a base set of entities and then merging all the sets of information to provide a focused view of the data."	L03	187	Theme
verification - "Verification authenticates, corrects, standardizes, and augments records against an external standard most often a database. For example, most companies standardize customer files against the United States Postal Service database."	EC02	28	Synonym
data enrichment - "is normally the product of data integration will occur when an additional attribute can be assigned to a data entity. For example, if external data is being introduced to the data warehouse, the data entity 'Customer' might be enriched by a new attribute, called C1, which was culled from an econometric source database."	KE95	139	Synonym

Match and Consolidate Data

match and consolidate data - "examines data to find duplicate records for a single real-world entity such as Customer or Product, both within a single database or file and across different files, and then consolidates the data into single occurrences of records."	EN99	262	Theme
matching, or deduplication - "involves the elimination of duplicate standardized records."	KI04	156	Synonym
matching - "Matching identifies records that represent the same individual, company, or entity. Vendors offer multiple matching algorithms and allow users to select which algorithms to use on each field."	EC02	28	Synonym
survivorship - "refers to the process of distilling a set of matched (deduplicated) records into a unified image that combines the highest-quality column values from each of the matched records to build conformed dimension records."	KI04	158	Theme
consolidation - "is a catchall term for those processes that make use of collected metadata and knowledge to eliminate duplicate entities and merge data from multiple sources, among other data enhancement operations."	L03	152	Theme
consolidation/householding - "Consolidation combines the elements of matching records into on complete record. Consolidation also is used to identify links between customers, such as individuals who live in the same household, or companies that belong to the same parent."	EC02	28	Synonym
householding - "is a process of reducing a number of records into a single set associated with a single household."	L03	156	Synonym

customer matching and householding - "Combining data about customers from disparate data sources is a classic data warehousing problem. It may go under the names of de-duplicating or customer matching, where the same customer is represented in two customer records because it hasn't been recognized that they are the same customer. This problem may also go under the name of householding, where multiple individuals who are members of the same economic unit need to be recognized and matched."	KI98	302	Synonym
elimination of duplicates - "is a process of finding multiple representations of the same entity with the data set and eliminating all but one of those representations from the set."	L03	156	Synonym
merge/purge - "involves the aggregation of multiple data sets followed by eliminating duplicates."	L03	156	Theme
Filter - "Filtering involves deleting duplicate, missing, or nonsensical data elements, such as when an ETL process loads the wrong file or the source system corrupts a field. Caution must be taken when filtering data because it may create data integrity problems."	EC02	23	Theme

Analyze Data Defect Types

analyze data defect types - "This step analyzes the patterns of data errors for input to process improvements."	EN99	265	Theme
Detect and Report - "In some cases, you may not want to change defective data because it is not cost-effective or possible to do so…In these cases, analysts need to notify users and document the condition in meta data."	EC02	23	Synonym

Prevent Future Errors

Prevent - "Prevention involves educating data entry people, changing or applying new validations to operational systems, updating outdated codes, redesigning systems and models, or changing business rules and processes."	EC02	23	Theme
data edits - "are computerized routines that verify whether data values and their representations satisfy prespecified constraints…Data editing capabilities are built into modern database management systems, although editing may be conducted elsewhere."	RE92	246	Theme
data edits - "computerized routines, which verify whether data values and/or their representations satisfy predetermined constraints."	RE96	23	Synonym
data editing - "cleaning up a small portion of the data each day…cleaning up the new data created daily and cleaning up the data before they are used."	RE01	55	Synonym
edit controls - "involve business rules based on the domains of data values permitted for a given field, pair of fields, and so on."	RE01	119	Synonym
error checks - "A series of data-quality screens or error checks are queued for running - the rules for which are defined in metadata."	KI04	136	Synonym

Data Cleansing Tools

data cleansing tools - "are designed to examine data that exists to find data errors and fix them. To find an error, you need rules. Once an error is found, either it can cause rejection of the data (usually the entire data object) or it can be fixed. To fix an error, there are only two possibilities: substitution of a synonym or correlation through lookup tables."	OL03	21	Theme
data reengineering, cleansing, and transformation tools - "Data 'correction' tools that extract, standardize, transform, correct (where possible), and enhance data, either in place of or in preparation for migrating the data into a data warehouse."	EN99	312	Theme

Data Cleansing Tool Functionality

data cleansing tools strengths - "Data cleansing companies provide support for processing selective data fields to standardize values, find errors, and make corrections through external correlation. Their target has been primarily name and address field data, which easily lends itself to this process. It has also been found to be usable on some other types of data."	OL03	53	Theme
data quality tool core capabilities - "Data Auditing…Parsing…Standardization…Verification…Matching…Consolidation/Householding"	EC02	27-28	Theme
customer-centric data quality tools - "Traditionally, vendors have focused on name and address elements because they are the most volatile fields in corporate databases…they have developed robust parsing engines and extensive reference libraries to aid in standardizing data, and build sophisticated algorithms for matching and householding customer records."	EC02	27	Theme

data cleansing appropriateness - "Cleansing data is often used between primary databases and derivative databases that have less tolerance for inaccuracies…Data cleansing has been specifically useful for cleaning up name and address information. These types of fields tend to have the highest error rate at capture and the highest decay rates, but also are the easiest to detect inaccuracies within and the easiest to correct programmatically."	OL03	97	Theme
data cleansing tool functionality - "deal with INVALID values in single data elements or correlation across multiple data elements. Many products are available to help you construct data cleansing routines."	OL03	59	Theme
clean-up tools - "the biggest issues are developing the business rules and ensuring that the tool scales to the size of the clean-up effort. Some clean-up tools are of the general-purpose variety, allowing the user to define his or her domains of allowed data values. Others, such as those based on the Postal Standard, come fully equipped with rules."	RE01	119	Theme

Emerging Data Cleansing Tool Functionality

data quality tool emerging capabilities - "Non-name and Address Data…Internationalization…Data Augmentation…Real-Time Cleaning…Customer Key Managers…Integration With Other Tools…Data Integration Hubs"	EC02	30-31	Theme
non-name and address data - "Vendors are developing parsing algorithms to identify new data types, such as emails, documents, and product numbers and descriptions. They are also leveraging standardization and matching algorithms to work with other data types besides names and addresses."	EC02	20	Theme
internationalization - "To meet the needs of global customers, vendors are adding support for multi-byte and unicode character strings. They are also earning postal certifications from the U.S., Canada, Australia, and Great Britain, and adapting to address reference files in other countries."	EC02	30	Theme
data augmentation - "While the USPS database can add zip+4 and other fields to a record, some vendors now can augment addresses with geocode data (I.e. latitude/longitude, census tracts, and census blocks) and demographic, credit history, and psychographic data from large information service providers such as Polk, Equifax, and Claritas."	EC02	30	Theme
geographic enhancement - "data enhanced with geographic information allows for analysis based on regional clustering and data inference based on predefined geodemographics. The first kind of geographic enhancement is the process of address standardization, where addresses are cleansed and then modified to fit a predefined postal standard, such as the United States Postal Standard. Once the addresses have been standardized, other geographic information can be added, such as locality coding, neighborhood mapping, latitude/longitude pairs, and other kinds of regional codes."	L03	191	Synonym
demographic enhancement - "Demographics describe the similarities that exist within an entity cluster, such as customer age, marital status, gender, income, and ethnic coding…Demographic enhancements can be added as a by-product of geographic enhancements or through direct information merging."	L03	191	Synonym
psychographic enhancement - "Psychographics describe what distinguishes individual entities within a cluster. For example, psychographic information can be used to segment the population by component lifestyles, based on individual behavior…The trick to using psychographic data is in being able to make the linkage between the entity within the organization database and the supplied psychographic data set."	L03	192	Synonym
customer key managers - "Some vendors are marketing internal match keys as a convenient way to associate and track customers across time and systems."	EC02	31	Theme
integration with other tools - "Many vendors offer a software developer's kit (SDK) which makes it easy for ETL and application vendors to embed data cleansing routines into their applications."	EC02	31	Theme
data integration hubs - "Data integration hubs channel <disparate system> interfaces into a central repository that maps incoming data against a clean set of standardized records."	EC02	31	Theme
real-time dimension manager system - "used primarily on customer information, converts incoming customer records, which may be incomplete, inaccurate, or redundant, into conformed customer records…typically modularized into the following subcomponents: * Cleaning...* Conforming...* Matching...* Survivorship...* Publication"	KI04	447-451	Synonym

real-time cleaning - "Traditionally, data quality tools clean up flat files in batch on the same platform as the tool. Most vendors now offer tools with a client/server architecture so that validation, standardization, and matching can happen in real time across a local-area network or the Web."	EC02	30	Synonym

Table B-3: Information Stewardship Content Analysis

Reference Key	
BR00	Brackett (2000)
BR94	Brackett (1994)
BR96	Brackett (1996)
EC02	Eckerson (2002)
EN99	English (1999)
HU99	Huang, et al. (1999)
KE95	Kelly (1995)
KI04	Kimball & Caserta (2004)
KI98	Kimball, et al. (1998)
LO03	Loshin (2003)
OL03	Olson (2003)
RE01	Redman (2001)
RE92	Redman (1992)
RE96	Redman (1996)

Description	***Reference***	***Page Number***	***Theme/Synonym***
Information Stewardship Defined			
information stewardship - "is 'the willingness to be accountable for a set of business information for the well-being of the larger organization by operating in service, rather than in control of those around us.'"	EN99	402	Theme
data resource management - "is the business activity responsible for designing, building, and maintain the data resource of the organization and making data readily available for developing information…It is an enormous task to refine data, remove redundancies, identify variability and designate official data variations, and develop a formal data resource while continuing to support business operations. The task requires a chief data architect supported by a staff of data architects and data engineers to face the challenges and build a common data architecture."	BR92	29	Synonym
data engineering - "is the discipline that designs, builds, and maintains the data resource library…"	BR96	48	Theme
data engineering - "It is a discovery process that relies largely on people to determine the true meaning of disparate data. It takes real thought, analysis, intuition, and consensus by knowledgeable people to identify the true content and meaning of disparate data."	BR92	13	Synonym

information engineering - "is the discipline for identifying information needs and developing information systems that produce messages that provide information to a recipient."	BR96	44	Synonym
information stewardship objectives - "business accountability for information quality…business 'ownership' of data definition…data conflict resolution mechanism…improve business and information systems partnership"	EN99	403	Theme
business accountability for information quality - "Improve the value and quality of information, and decrease the costs of nonquality information"	EN99	403	Theme
business 'ownership' of data definition - "Increase business communication, understanding and productivity through data as a common business language"	EN99	403	Theme
data conflict resolution mechanism - "Maximize data value through quality shared data with common definition, and minimize data costs through eliminated nonshared or redundant databases, interfaces, and applications"	EN99	403	Theme
improve business and information systems partnership - "Improve customer satisfaction and team effectiveness"	EN99	403	Theme
consensus - "Consensus is the best approach to developing a common data architecture and refining disparate data."	BR92	181	Synonym
facilitated approach - "A consensus approach to refining data that involves a group of knowledgeable people requires facilitation to ensure that consensus is reached."	BR92	183	Synonym

Data Quality Program

data quality program components - "* Clear business direction, objectives, and goals; * Management infrastructure…; * An operational plan…; * Program administration"	RE96	18-19	Theme
data quality system - "By the phrase 'data quality system (DQS),' we mean the totality of an organization's efforts that bear on data quality."	RE01	75	Synonym
information quality program - "To establish an information quality program, the information product manager can adapt classical TQM principles…Adapting the TQM literature, five tasks should be undertaken: Articulate an IQ Vision in Business Terms…Establish Central Responsibility for IQ Within through the IPM...Educate Information Product Suppliers, Manufacturers, and Consumers...Teach New IQ Skills...Institutionalize Continuous IQ Improvement."	HU99	27-28	Synonym
data quality assurance program - "For companies to create high-quality databases and maintain them at a high level, they must build the concept of data quality assurance into all of their data management practices. Many corporations are doing this today and many more will be doing so in the next few years. Some corporations approach this cautiously through a series of pilot projects, whereas some plunge in a institute a widespread program from the beginning."	OL03	65	Synonym
data quality assurance initiatives - "are becoming more popular as organizations are realizing the impact that improving quality can have on the bottom line."	OL03	23	Synonym
data stewardship program - "The best way to kickstart a data quality initiative is to fold it into a corporate data stewardship or data administration program."	EC02	15	Synonym
data quality assurance activities - "There are three primary roles the group can adopt…One of them, project services, involves working directly with other departments on projects. Another, stand-alone assessments, involves performing assessments entirely within the data quality assurance group. Both of these involve performing extensive analysis of data and creating and resolving issues. The other activity, teach and preach, involves educating and encouraging employees in other groups to perform data auditing functions and to employ best practices in designing and implementing new systems."	OL03	75	Theme
data quality assurance program for data accuracy - "The assertion is that any effective data quality assurance program includes a strong component to deal with data inaccuracies. This means that those in the program will be looking at a lot of data."	OL03	65	Theme
management procedures - "reasonable management procedures <for data resources> must be rigorous and reasonable…Adequate data responsibility includes centralized control of the data resource architecture."	BR00	217-218	Theme

data quality assurance methods - "The inside-out method starts with analyzing the data. A rigorous examination using data profiling technology is performed over an existing database. Data inaccuracies are produced from the process that are then analyzed together to generate a set of data issues for subsequent resolution...<Outside-in> method looks for issues in the business, not the data. It identifies facts that suggest that data quality problems are having an impact on the business...These facts are then examined to determine the degree of culpability attributable to defects in the data."	OL03	73	Theme
data quality project plan - "prioritizing projects that have the greatest upside for the company, and tackle them one by one."	EC02	16	Theme
Data Quality Policy			
data quality policy - "A statement of management's intent regarding data and information quality, the organization's long-term data and information quality improvement objectives, and specific management accountabilities for pursuing the intent and achieving the objectives. The policy is intended as a 'guide for managerial action'."	RE01	80	Theme
data policy - "enterprises desirous of improving data quality and getting full benefit from data can and should establish clear management responsibilities for data. Based on the issues it faces and its deployment capabilities, an enterprise should consider a data policy that covers the following areas. * Quality in its broadest sense; * Data inventory; * Data sharing and availability; * Data architecture; * Security, privacy, and rules of use; * Planning.	RE96	52	Synonym
Information Stewardship Guidelines			
support tools for information stewards - "information policy…training…information stewardship guidelines"	EN99	417	Theme
information stewardship guidelines - "Topics should include an introduction to the definition and purpose of stewardship, role and responsibility descriptions, support resources available, guidelines for data definition, information quality standard setting, data access clarifications, and other tasks."	EN99	417	Theme
Executive Support for a Data Quality Program			
executive buyin to launch a data quality program - "To succeed, a data quality program must be initiated by the CEO, overseen by the board of directors, and managed either by a chief data quality officer or senior-level business managers in each area of the business."	EC02	15	Theme
senior management critical to data quality program success - "After the prototype stage, programs move further and faster with senior leadership. No enterprise can hope to build data quality into its mainstream without it."	RE96	66	Theme
senior management critical to data quality program success - "There is no question that leadership of senior management is critical to the long-term success of quality programs. This is particularly true in data quality…Senior management should promote a value structure within the enterprise so process owners act in the enterprise's interests. Management must also ensure that owners of critical processes and data keepers are in place and that they have needed authority and resources to do their jobs."	RE92	261	Theme
CEO critical to data quality success - "An organization's most senior leader must not delegate responsibility for data quality."	RE01	5	Theme
Business Customer Involvement in the Data Quality Program			
knowledge people must develop common metadata - "The best way to develop good common metadata is to include business experts, domain experts, and data experts in the development effort. The business experts know the specific business rules and processes unique to the organization or organizations within the scope of the common metadata. The domain experts know the discipline involved in the common metadata, such as water resources, health care, surveying, and land use. The data experts know how data are managed from the real world through logical design to physical implementation."	BR96	193	Theme
business-oriented information engineering - "A business understanding requires direct client involvement - the direct involvement of people knowledgeable about the business and the data supporting the business. The best approach to building a common data architecture is a partnership between data architects, data engineers, and knowledgeable clients. The partnership allows clients to exploit their knowledge of the business and the data supporting the business to build a common data architecture."	BR92	37-38	Theme

direct client involvement in data integrity - "Data architects will design and maintain the common data architecture and build the formal data resource, but clients will use that architecture and populate the data resource to support their business activities. Defining data integrity in an understandable way helps clients become involved in defining and implementing data integrity."	BR92	149	Theme
direct client involvement in data documentation - "Good data documentation requires client involvement. Clients generally have a better knowledge and understanding of the business and data that support the business than the data processing staff."	BR92	154	Theme
data definition team - "The most effective way to establish common and consensus data definition is to conduct facilitated data definition sessions involving representatives of all business areas that have a stake in a business subject or common collection of information."	EN99	413	Theme
direct client involvement in data definitions - "One excellent way to develop data definitions that are meaningful to the business is to include business clients in the preparation of those data definitions."	BR00	64	Theme
direct client involvement in a data resource quality initiative - "Another good practice is to ensure the direct involvement of knowledgeable business clients in a data resource quality initiative. The successful initiatives that I have seen involve a mix of business clients and technical staff."	BR00	258	Theme
Information Stewardship Team			
information stewardship teams - "There are two key stewardship teams: the business information stewardship team and the executive information steering team."	EN99	413	Theme
Data Quality Council			
data quality council - "The senior management body charged with executing the data quality policy at the highest level."	RE01	79	Theme
executive information steering team - "either appoints business information stewards or gives authority to the selected stewards to carry out the responsibilities they have. This authority includes making the time available from the steward's schedules..."	EN99	413	Synonym
corporate stewardship committee - "needs to develop a master plan for data quality that contains a mission statement, objectives, and goals. It then needs to educate all employees about the plan and their roles in achieving the goals…The corporate stewardship committee also needs to oversee and provide direction to all data quality teams or functions scattered throughout the company."	EC02	15	Synonym
data quality assurance advisory group - "The data quality assurance team must decide how it will engage the corporation to bring about improvements and return value for their efforts. The group should set an explicit set of guidelines for what activities they engage in and the criteria for deciding one over the other. This is best done with the advisory group."	OL03	75	Synonym
Data Quality Team			
data quality assurance department - "This should be organized so that the members are fully dedicated to the task of improving and maintaining higher levels of data quality. It should not have members who are part-time. Staff members assigned to this function need to become experts in the concepts and tools used to identify and correct quality problems."	OL03	69	Synonym
business information stewardship team - "provide the business validation for data definition."	EN99	413	Synonym
Data Quality Team Job Functions			
data quality team roles - "Chief Quality Officer…Data Steward…Subject Matter Expert…Data Quality Leader…Data Quality Analyst…Tools Specialist…Process Improvement Facilitator…Data Quality Trainer"	EC02	17	Theme
information quality job functions - "information quality manager or leader…information architecture quality analyst…data cleanup coordinator, data quality coordinator, or data warehouse quality coordinator…information quality analyst…information quality process improvement facilitator...information quality training coordinator"	EN99	451-453	Theme
Chief Quality Officer			
Chief Quality Officer - "A business executive who oversees the organization's data stewardship, data administration, and data quality programs."	EC02	17	Theme

Strategic Data Steward

strategic data steward - "is a person who has legal and financial responsibility for a major segment of the data resource. That person has decision-making authority for setting directions and committing resources for that segment of the data resource. The strategic data steward is usually an executive or upper-level manager and usually has responsibility along organizational lines, much as the director of human resource is the strategic data steward for human resource data."	BR00	213	Theme
Data Quality Leader - "Oversees a data quality program that involves building awareness, developing assessments, establishing service level agreements, cleaning and monitoring data, and training technical staff."	EC02	17	Synonym
information quality manager - "is accountable for implementing processes to assure and improve information quality."	EN99	451	Synonym

Tactical Data Steward

tactical data stewards - in very large organizations, "the best approach is to designate tactical data stewards between the strategic data stewards and detail data stewards to manage the international aspects of the data resource."	BR00	217	Theme

Detail Data Steward

detail data steward - "is a person who is knowledgeable about the data by reason of having intimate familiarity with the data. That person is usually a knowledgeable worker who has been directly involved with the data for a considerable period of time. The detail data steward is responsible for developing the data architecture and the data resource data. That person has no decision making authority for setting directions for the data resource or committing resources to data resource development."	BR00	214	Theme
information architecture quality analyst - "is responsible for analyzing and assuring quality of the data definition and data model processes."	EN99	451	Synonym
information steward - "is accountable for defining the information strategy. This person formalizes the definition of analytic goals, selects appropriate data sources, sets information generation policies, organizes and publishes metadata, and documents limitations of appropriate use."	KI04	118	Synonym
data steward - "is a person who watches over the data is responsible for the welfare of the data resource and its support of the business, particularly when the risks are high. There are many terms that could be used, such as data guardians, data custodians, data coordinators, data analysts, data trustees, data curators, data administrators, data facilitators, data negotiators, data interventionists, and so on."	BR00	212	Synonym
data steward - "sometimes called the data administrator, is responsible for gaining organizational agreement on common definitions for conformed warehouse dimensions and facts, and publishing and reinforcing these definitions. This role is often also responsible for developing the warehouse's metadata management system."	KI98	70-71	Synonym
Data Steward - "A business person who is accountable for the quality of data in a given subject area."	EC02	17	Synonym
Data custodians, data stewards, or data trustees - "can be designated to coordinate policy accountabilities for the most important enterprise data."	RE96	51	Synonym
information product manager - "Companies should appoint an information product manager to manage their information processes and resulting products."	HU99	20	Synonym
information product manager's key responsibility - "is to coordinate and manage the three major stakeholder groups: the supplier of raw information, the producer or manufacturer of the deliverable information, and the consumer of the information. To do so, the information product manager must apply an integrated, cross-functional management approach. The information product manager orchestrates and directs the information production process during the product's life cycle in order to deliver quality information to the consumer."	HU99	25	Theme
information product manager defines IQ metrics - "the Information Product Manager (IPM) must develop the corresponding IQ metrics, upon defining IQ dimensions, to measure and analyze the quality of the information product and improve it accordingly."	HU99	59	Theme

Data Cleanup Coordinator

data cleanup coordinator - "is responsible for overseeing the data acquisition and cleansing activities of a data warehousing initiative, conversion, or cleanup initiatives."	EN99	452	Theme

information-quality leader - "detects, corrects, and analyzes data-quality issues."	KI04	118	Synonym
Tools Specialists - "Individuals who understand either ETL or data quality tools or both and can translate business requirements into rules that these systems implement."	EC02	17	Synonym
data keeper - "the data keeper's job is to care for data on behalf of the enterprise…A data keeper should be assigned to each database and has three explicit functions: * to ensure communication between users and creators of data. In this function, the data keeper ensures both that a single, consistent set of data quality requirements is used and that adequate feedback channels exist and are operational, * to manage the edits and their operation, and * to conduct any database cleanups, should they be needed."	RE92	241	Synonym
data keeper should maintain the data dictionary - "The data keeper should maintain a comprehensive data dictionary, which should provide definitions of stored data fields. In addition, it should provide definitions of all data fields in processes upstream of the database and the changes to these fields downstream."	RE92	244	Theme
Information Quality Analyst			
information quality analyst - "is responsible for assessing and measuring information quality and providing feedback."	EN99	452	Theme
data warehouse quality assurance analyst - "ensures that the data loaded into the warehouse is accurate. This person identifies potential data errors and drives them to resolution."	KI98	71	Synonym
data-quality specialist - "primarily works with the systems analyst and the ETL architect to ensure that business rules and data definitions are propagated throughout the ETL processes."	KI04	396	Synonym
Data Quality Analyst - "Responsible for auditing, monitoring, and measuring data quality on a daily basis, and recommending actions for correcting and preventing errors and defects."	EC02	17	Synonym
Information Quality Process Improvement Facilitator			
information quality process improvement facilitator - "facilitates improvements in information processes."	EN99	453	Theme
Process Improvement Facilitator - "Coordinates efforts to analyze and reengineer business processes to streamline data collection, exchange, and management, and improve data quality."	EC02	17	Theme
Information Quality Training Coordinator			
information quality training coordinator - "is responsible for overseeing the development and delivery of education, training, or awareness raising in information quality to all levels of personnel in the enterprise."	EN99	453	Theme
Data Quality Trainer - "Develops and delivers data quality education, training, and awareness programs."	EC02	17	Theme
Subject Matter Expert			
Subject Matter Expert - "A business analyst whose knowledge of the business and systems is critical to understand data, define rules, identify errors, and set thresholds for acceptable levels of data quality."	EC02	17	Theme

References

Beal, Barney. (March 9, 2005). “Report: Half of data warehouses to fail”. *SearchCRM.Com [Online]*. Retrieved September 13, 2005 from http://searchcrm.techtarget.com/originalContent/0,289142,sid11_gci1066086,00.html.

Brackett, Michael H. (2000). *Data Resource Quality: Turning Bad Habits into Good Practices*. Upper Saddle River, NJ: Addison-Wesley.

Brackett, Michael H. (1994). *Data Sharing*. New York: John Wiley & Sons, Inc.

Brackett, Michael H. (1996). *The Data Warehouse Challenge: Taming Data Chaos*. New York: Wiley Computer Publishing.

CSU Writing Lab (Colorado State University). (2005). “Conducting Content Analysis”. *Writing Guides*. Retrieved September 26, 2005 from http://writing.colostate.edu/guides/research/content/index.cfm.

Eckerson, Wayne W. (2002). “Data Quality and the Bottom Line”. *TDWI [Online]*. Retrieved September 13, 2005 from http://www.tdwi.org/research/display.aspx?ID=6028.

English, Larry P. (1999). *Improving Data Warehouse and Business Information Quality*. New York: Wiley Computer Publishing.

Faden, Mike. (April 10, 2000). “Data Cleansing Helps E-Businesses Run More Efficiently”. *InformationWeek [Online]*. Retrieved September 13, 2005 from http://www.informationweek.com/781/clean.htm.

Galliers, Robert D., & Baets, Walter R.J. (1998). *Information Technology and Organizational Transformation: Innovation for the 21st Century Organization*. New York: John Wiley & Sons.

Huang, Kuan-Tse, Lee, Yang W., & Wang, Richard Y. (1999). *Quality Information and Knowledge*. Upper Saddle River, NJ: Prentice-Hall.

Hudicka, Joseph. (March 20, 2003). “Bumpy Ride – Data Migration Projects Still Plagued by Problems”. *Intelligent Enterprise*, 10.

Kelly, Sean. (1995). *Data Warehousing: The Route to Mass Customization*. New York: John Wiley & Sons.

Kimball, Ralph, & Caserta, Joe. (2004). *The Data Warehouse ETL Toolkit*. New York: Wiley Computer Publishing.

Kimball, Ralph, Reeves, Laura, Ross, Margy, & Thornhwaite, Warren. (1998). *The Data Warehouse Lifecycle Toolkit*. New York: Wiley Computer Publishing.

Krippendorff, Klaus. (2004). *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: Sage Publications, Inc.

Leedy, Paul D, & Ormrod, Jeanne Ellis (2001). *Practical Research : Planning and Design*. New Jersey: Merrill Prentice Hall.

Loshin, David. (2003). *Business Intelligence*. San Francisco: Morgan Kaufmann Publishers.

Olson, Jack E. (2003). *Data Quality: The Accuracy Dimension*. San Francisco: Morgan Kaufmann Publishers.

Pendse, Nigel. (March 22, 2005). “Market share analysis”. *OLAP Report [Online]*. Retrieved September 13, 2005 from http://www.olapreport.com/Market.htm.

Redman, Thomas C. (1996). *Data Quality for the Information Age*. Boston: Artech House.

Redman, Thomas C. (1992). *Data Quality Management and Technology*. New York: Bantam Books.

Redman, Thomas C. (2001). *Data Quality: The Field Guide*. Boston: Digital Press.

St Clair, Guy. (1997). *Total Quality Management in Information Services*. London: Bowker Saur.

Bibliography

Atkins, Mark E. (May 3, 2000). “The Time Has Come for Enterprise Data Quality Management”. *Adviser Zone [Online]*. Retrieved September 13, 2005 from http://doc.advisor.com/doc/06456.

Bank Marketing Association. (April 2003). “10 Rules for Cleaner Customer Data”. *Bank Marketing, 15397890, Vol. 35, Issue 3,* 9.

Dodge, Gary, & Gorman, Tim. (2000). *Essential Oracle8i Data Warehousing*. New York: Wiley Computer Publishing.

Ericson, Jim. (February 7, 2002). “The Cost of Poor Data”. *E-Business News [Online]*. Retrieved September 13, 2005 from http://line56.com/articles/default.asp?articleID=3364&TopicID=3.

Fuld, Leonard M. (June 15, 1999). “Data Waste”. *CIO [Online]*. Retrieved September 13, 2005 from http://www.cio.com/archive/enterprise/061599_ic.html.

Glance, Kerry. (February 23, 2005). "Expert: Data quality is misunderstood". *SearchCRM.Com [Online]*. Retrieved September 13, 2005 from http://searchcrm.techtarget.com/qna/0,289202,sid11_gci1061156,00.html.

Goldenberg, Barton J. (2002). *CRM Automation.* Upper Saddle River, NJ: Pearson Prentice Hall.

Kotler, Philip, & Keller, Kevin. (2005). *Marketing Management 12e.* Upper Saddle River, NJ: Pearson Prentice Hall.

Luftman, Jerry N. (2004). *Managing the Information Technology Resource: Leadership in the Information Age*. Upper Saddle River, NJ: Pearson Prentice Hall.

Kirsner, Scott. (June 2001). "Information Waste". *Darwin [Online]*. Retrieved September 13, 2005 from http://www.darwinmag.com/read/060101/ecosystem.html.

MAS Strategies & Group1 Software. (2004). "Data Quality and Data Integration: The Keys for Successful Data Warehousing". *Group1 Software [Online]*. Retrieved September 13, 2005 from http://www.g1.com/Ad/SuccessfulDI/.

Palanisamy, Ramaraj. (September 2001). "Empirically Testing the Relationships Between User Involvement, Information Waste, and MIS Success". *Journal of Services Research, Apr-Sep2001, Vol. 1 Issue 1,* 70-103.

Redman, Thomas C. (May 24, 2005). "Data quality: Beware the dirty things your customers see". *SearchCRM.Com [Online]*. Retrieved September 13, 2005 from http://searchcrm.techtarget.com/columnItem/0,294698,sid11_gci1091536,00.html.

Turban, Efraim. (1995). *Decision Support and Expert Systems*. Englewood Cliffs, NJ: Prentice-Hall.

Turek, Norbert. (June 16, 2003). "Avoid Bad-Data Potholes". *InformationWeek, 944*, 51.