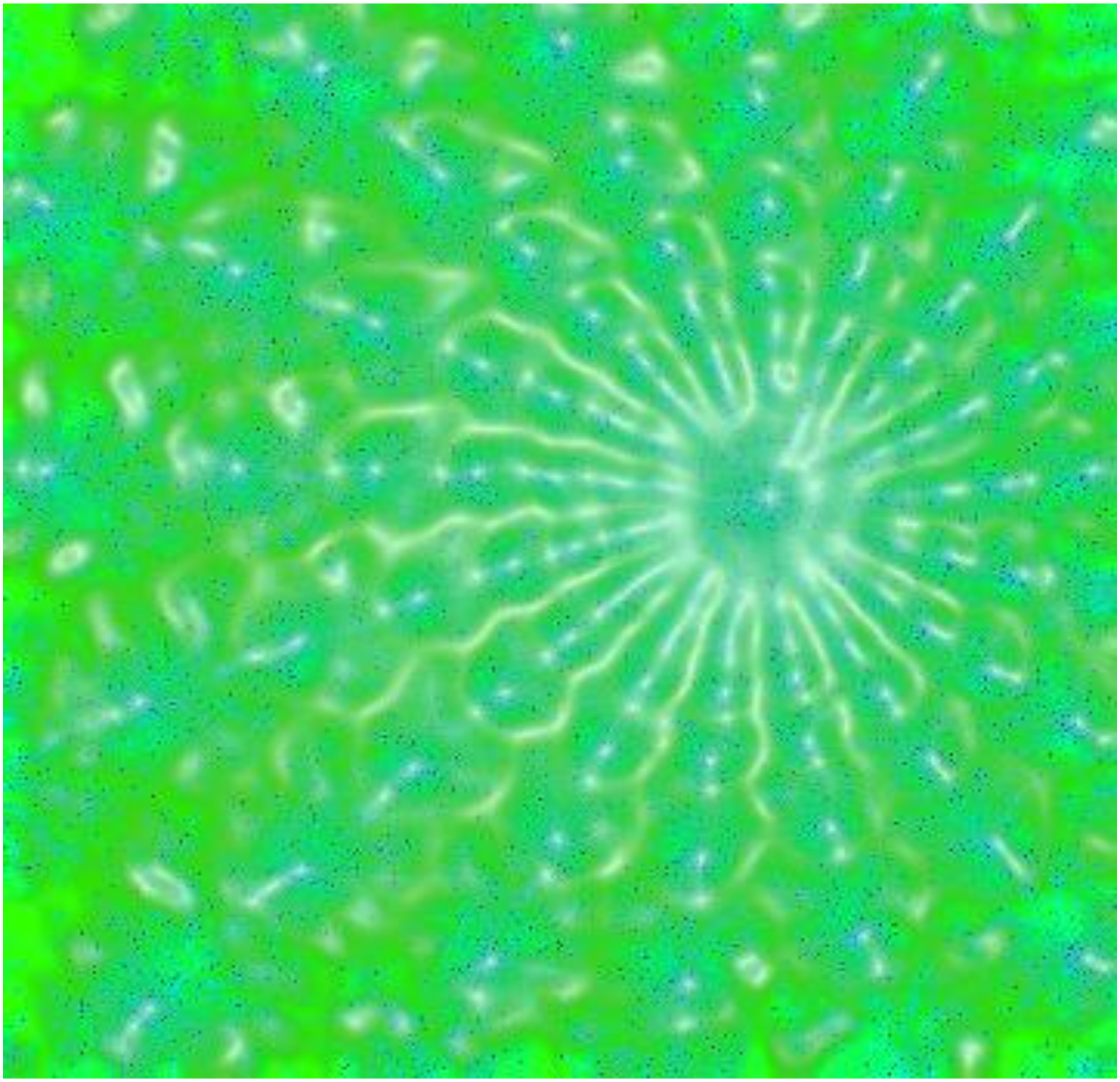
AN APPROACH FOR MANAGING DATA QUALITY

A thesis submitted in partial fulfillment of the requirements of the degree of
Master of Information Sciences (Research)

Charles Broderick Palmer
Faculty of Information Sciences and Engineering
University of Canberra, Australia
November 2011



ACKNOWLEDGMENTS

This thesis would not have been designed and written without the support, understanding (and endless cups of tea) from my ever-patient wife Liana who considered this work *'my other wife'.*

My appreciation for the panel that bore the ever-present risk that a new thesis author may not complete:

- Professor Craig McDonald (Panel Chair),
- Professor John Campbell (ADR)
- Ass. Professor Richard Lucas (HoD)

They all offered guidance, suggestions and encouragement throughout the journey.

I am indebted to my faculty and workplace colleagues who proofed, discussed software design issues and offered many suggestions and ideas that have helped to bring this work to fruition.

Special thanks to Annette Vincent, Neil Lynch and Dale Mackrell.

Finally the examiners who clearly spent a considerable amount of time and effort reading this work and offering constructive criticism whilst noting that the work is innovative and opens new areas for research in the data quality management domain.

ABSTRACT

Organisations manage data that supports decision-making activities. As data storage costs continue to fall and organisational appetite for more and persistent information expands, so the problems created by poor or variable data become more pervasive.

Financial, operational, social and legal issues associated with poor or inappropriate decision making are extensively documented.   However many organisations fail to manage their data quality issues effectively; or even at all. Data quality management is costly, particularly when much of the effort is directed to non-critical data.

This thesis reports on research that developed a method to better target data quality effort and built a software artefact to explore the validity of the method.

The method is to identify, rank and sort data using a mix of technical, user and business-based ranking points to reflect the usage and importance value for each data element.

The software artefact and method were tested in an experimental setting where different levels of random quality errors were introduced into a database and the impact of the errors assessed. The relative merits of quality assurance using ranked data elements rather than unranked elements have been demonstrated.

The research was based on an extensive review of academic literature, commercial literature, and current commercial data quality products and services.

The thesis demonstrates the validity of a ranking approach to data quality. The method provides a means for organisations to improve their data quality assurance and thereby to improve their decision making confidence.

This research makes a significant contribution to the principles of managing data quality.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF TRADEMARKS AND LOGOS

Microsoft trademarks, products and logos have been used in accordance with the directives noted at: http://www.microsoft.com/about/legal/en/us/IntellectualProperty/Trademarks/Usage/General.aspx, Viewed 26 October 2011

Oracle	-	Oracle™ RDMS
IBM™	-	International Business Machines
MySQL	-	Open Source RDMS
ODBC	-	Open Data Base Connection
JET™	-	Microsoft's MS ACCESS™ data binding services

1 AN APPROACH FOR MANAGING DATA QUALITY

1.1 INTRODUCTION

The need for organisations to identify, measure and manage the quality of their data is becoming increasingly vital as technology becomes more complex and organisations experience expanded information and decision demands from political, social, economic and technical influences.

An example of a decision based on flawed data has been documented and tabled in the House of Commons, 2002-2003, (Vol.1, Para 1-17). These documents describe poor intelligence data and the consequential decision to embark on the invasion of Iraq. The US costs estimated in human lives and funding vary substantially and range from US$801 billion (Belasco, 2011) to US$3 trillion (Stiglitz & Bilmes, 2010).

Private and public organisations rely upon information to make decisions and, as well, are becoming more accountable for their IT investments, business decisions and achievement of performance goals.

The complexity presented by information quality management is that many users can assign different quality measures and definitions to a common information item at the same time. As this information retains some level of validity in an organisation over time, so too the mix of quality levels, and quality measures remain.

Managing quality against data elements can become financially and logistically overwhelming in many organisations.

There are two broad business costs associated with data quality management:

1. The costs associated with the definition, detection and correction of data quality issues. These activities are expensive and may offer a poor yield for investment. It is unlikely that the time and costs associated with poor data quality measurement would be different between high and low value data; and

2. The costs associated with discovery of data quality issues during operations. This discovery process can present high direct and indirect operational costs.

Are these costs the reason that few organisations have a plan for managing data quality issues as a strategic and production-focused direction?

In summary, there are significant risks associated with poor data quality:

1. The first cost (definition, detection and correction) is the result of planned management;

2. Risk costs associated with detection-by-disaster and recovery;

3. The cost of poor decision making, re-work and potential loss of business;

4. The costs associated with detecting data quality errors and then applying remedial action against low value data; and

5. Potential for low trust-levels within the organisation based on assumed poor data quality.

Perhaps the costs associated with detection are considered too high when compared to detection-by-disaster?

This thesis shows how data quality could be managed by ranking data into a high-value data set and then examining quality issues with this set rather than attempting to address all data quality issues regardless of the data's value.

Organisations already rank some of their business decisions based on financial or business risks. A common example is where many organisations decide (or *should* decide) not to post debtor's reminder statements for amounts less than $1.00. The cost of generating and posting the statement as an attempted recovery effort outweighs the benefit should the recovery be successful. Using similar logic, the cost of examining low value data for quality errors may not be justified when compared to the benefits offered by detecting data quality issues against low value data. How then can an organisation calculate the value of their data in order to make a defensible judgment against the cost of measurement?

This thesis describes experiments that demonstrate a method that ranks data by valuing data items in terms of usage and frequency. This approach offers some key benefits to the data users:

1. Although the data quality error <u>count</u> detected and considered for remedial action is unlikely to change, the <u>value</u> of the data examined and considered for remedial action is significantly increased for the same level of effort and expense;

2. This approach demonstrates an opportunity where data analysis and detection can be scaled to reflect the organisation's data usage and appetite for data quality risks;

3. Given the potential economies that this approach offers, a recurring remediation plan to address data at some value level can offer a known business outcome;

4. Recurring analysis of high-value data offers a level of comfort to data users and decision makers; and

5. This approach also offers the opportunity to inform data users the percentage of data (and the scalar ranking of the untested data) they are using that has not been examined for errors.

Therefore the aim of this research is to devise and demonstrate a method that allows organisations to better deploy their data quality management resources more effectively and efficiently.

1.2 DATA QUALITY DEFINITIONS

These definitions have been derived from variety of sources and contexts.

Information and Data Quality. Rudra & Yeo, 1999, p2 state that information quality and data quality are not differentiated. In this thesis 'data' and 'information' are similarly not differentiated.

Data is a representation of some event or state.

Quality describes a measure describing an acceptable level of defects. Juran 1998, p2.3 describes quality as "*Quality means freedom from deficiencies*".

Data quality is a set of measures that describe a condition where the examined data exhibits some level of tolerable or intolerable defects "*data quality is not the absence of defects; it is the absence of intolerable defects.*" (McKnight & William, 2010, p3).

Data Quality Measures are defined as a set of protocols that provide a template and methods for measuring data quality. The Canadian Institute for Health Information (CIHI, 2007, pp1-10) describes data quality as a measure "*…appropriate to use for the purpose in question.*"

Data Quality Management (McKnight & William,2010,p3) "*Proper data quality management is also a value proposition that will ultimately fall short of perfection, yet provide more value than it costs*". Data quality management should then present a return on the investment required to conduct the activity.

Information Assurance (in this thesis) is treated as data quality assessment and management.

Data Quality Rules (DQR) refer to the definitions of acceptable data quality. DQR can be (and often is) specific to an organisation. Breached rules can cost time, money, customers, or in the case of medical databases, even present life threatening outcomes. (Yakout, Elmagarmid, & Neville, 2010, pp4-6).

1.3 THESIS STRUCTURE

This thesis describes the problem, research, motivation and a possible approach to the business issue surrounding data quality management.

This thesis is structured as seven chapters and appendices.

Ch.2 Literature Review

This chapter describes the literature review outcomes and classifies the results of the review into sections that describe the issues, current solutions and research that has been conducted in the data quality management domain is also examined. The sections in this chapter reflect some of the data quality domain facets.

Ch.3 The Research Problem

This chapter distills the research problems highlighted by the literature review. The problem areas identified are:

- Data Quality and Decision making
- Data Quality Frameworks
- The Changing Perceptions of Data Quality
- Data Quality in Supply Chains
- Cost and Value of Data Quality
- Ranking Approaches
- Current Solutions

Ch.4 Research Design

This chapter describes how a design science approach was used to analyse, design and develop an artifact (Ranking Tool) and to conduct experiments that test the planned ranked data approach.

Ch.5 Software Ranking Tool Design and Testing

Describes the design, development and testing of the Ranking Tool in five major iterations.

Ch.6 Experiment Findings

This chapter describes the conduct of the experiments using the ranking artifact and the findings that came from the experiments.

Ch.7 Conclusion

This chapter illustrates how the research problems from Chapter 3 were addressed in the research process and the associated experiments.
The contribution to the data quality management space is described.
Possible future research directions are noted.

2 LITERATURE REVIEW

Chapter Introduction

This chapter describes the research process that defines the data quality domain as well as costs and risks related to data quality. Research that had been conducted to rank information was also examined.

The outcome is that there has been much work describing and categorising data quality characteristics but little that describes the scoring and ranking process this thesis proposes.

This chapter has been classified into seven sections shown below. Each section describes the relevant research with context described at the end of each section:

1. **Data Quality and Decision making**
This section emphasises the link between information quality and decision making. The link is well established by many authors.

2. **Data Quality Frameworks**
This section describes a brief history of data quality frameworks that shows how the classification of data quality management has evolved to reflect multiple users using the same data in an organisation.

3. **The Changing Perceptions of Data Quality**
This section shows how data quality perceptions now include multiple quality measures for common data sets based on common and different users.

4. **Data Quality in Supply Chains**
An emerging trend where ‘just in time’ procurement and other related inventory management approaches are driving chained organisations who share dependencies for their part in production chains. An information quality issue can have profound repercussions throughout a set of chained corporations.

5. **Cost and Value of Data Quality**
The costs associated with data quality are outlined. Essentially the costs of detection may appear to outweigh the costs of remediation upon surprise detection. The essence is that if the costs of detection can be better managed, then perhaps the costs of prevention might become more apparent.

6. **Ranking Approaches**
Although there has been little literature that describes the ranking approach illustrated in this thesis, many authors have explored various ranking mechanisms that have been adapted in this thesis.

7. **Current Solutions**
There are many commercial solutions designed to detect and correct data quality issues. These approaches have many common features, but do not address all of the key quality issues associated with organisational data.

The approach described in this thesis complements rather than replaces these commercial approaches.

2.1 DATA QUALITY AND DECISION MAKING

Tayi & Ballou,1998, pp54-57 and Levis, Helfert & Brady (n.d.) both describe the strong relationship between information quality and effective and defensible decision making. O'Brien,2011,p1 notes "*high costs of low quality data and the cost of poor data quality …other serious consequential effects relating to tactical decision making and strategy generation*" when describing the imperatives for data governance and ownership.

Much of the literature concludes that the quality of data held in information systems is critical for organisational decision making.

Price & Shanks,2005(a),p1 state "*The effectiveness of an organization is dependent on the quality of its information…*" and "*Quality information and information quality management in an organization is essential for effective operations and decision-making*" (Price & Shanks,2005(b),p658). Information systems managers well understand that they need acceptable data quality as a key driver to the effectiveness of the decision making.

Seddon, Staples, Patnayakuni, et al., 1996, p165 comment that "*The value added by an organization's IT assets is a critical concern to both research and practice*". Given that data storage, retrieval and management are fundamental to many IT assets, information quality reflects a key yield of an IT asset. Seddon, Staples, & Patnayakuni,1999, p165 also note "*Total annual worldwide expenditure on information technology (IT) probably exceeds one trillion US dollars per year and is growing at about 10% compounded annually.*" Inappropriate information quality degrades the expected return on investments (ROI) in an IT environment.

Data holdings are used to present information that is then used to drive decision making for a range of business activities. Both public and private organisations are being held increasingly accountable to their shareholders and the public for accurate and informed decision making. A well publicised example is the Sarbanes-Oxley Act of 2002 that specifies particular information for US based organisations was ratified "...*to protect investors by improving the accuracy and reliability of corporate disclosures made pursuant to the securities laws, and for other purposes*".

Findings from the Australian National Audit Office (ANAO) note poor data quality as a key impediment to effective business decision making. The ANAO reported when reviewing the Australian Tax Office (ATO), ATO Data and Systems Quality,1998-99, p84 note *"that the quality of its main databases was probably somewhere between unsatisfactory and average...".* As a key revenue stream for the Australian Government, this highlights a significant issue.

ANAO, 2004, also reported that data quality was a key issue when reviewing The Health Insurance Commission (HIC) that *"…information quality management issues with specific references to timeliness, accuracy, accessibility, coherence, cost effectiveness and review as key information quality priorities."*

There are many causes for increased data range and volume increases. Sir Peter Gershon 2008, pp17-21 commented that *"Advances in processing power, storage and memory technologies have paved the way for more sophisticated use of data analytics and business intelligence technologies."* .

In addition to the technical advances, bespoke (customised or in-house developed) systems have become increasingly expensive to design, develop and maintain when compared to commercial systems "off-the-shelf" (COTS). Commercial systems have been designed to appeal to the widest possible audience to enhance sales opportunities and functionality to allow competition with other vendors as well as leveraging the development effort. For example, a typical COTS payroll system, chris^{21}™ (© Frontier Software Pty Limited, Victoria 2006) present access to approximately 7500 data columns whereas a similar system PERSPECT™ (Aspect Limited, ACT 1993) offered less than 3500 data columns. Additional columns allowed more history and a greater diversity of business information storage; as well as the propensity for related data quality errors.

Sir Peter Gershon's report (2008), p79, in his criticism of agency-level autonomy with ICT acquisition, stated: *"I also recommend strengthening governance regarding the adoption or modification of COTS/GOTS".* (Commercial Off-The-Shelf/Government Off-The-Shelf Systems). This comment mitigates the costs associated with software development 'from scratch' but does present a 'one size fits all' approach that offers larger and more complex data structures that a purpose-designed system thus increasing the range and detail of data that can be collected and managed.

Typical data holdings for organisations range from approximately 30,000 data elements (Centrelink) (personal conversation, 2007) to over 100,000 data elements for diverse activity organisations (ACT Government Analysts, 2006), (Skinner, 2009), (Tu Pham, 2007). Complex and diverse data holdings can better support complex data interrogation approaches and corresponding decision support.

Research can also be adversely affected by poor data quality as noted by Missier, Embury & Greenwood et al., (2006) , p977 where they concluded that data in public repositories may affect the validity of experimental results"*…and of the threat that poor data quality poses to the validity of experimental results."*. Given that research relies on appropriate information with which to draw conclusions, the 'flow on' effect of research based on flawed information could be significant.

Raden,2006,pp5-6 comments in a paper commissioned by Silver Creek Inc. that *"Service-Oriented Architectures will exacerbate the situation by further abstracting data from its source application …".* The emerging trend is to publish organisational data sets as web-based information. This trend, when presented using mash-ups and Web2 technologies, further removes the end-consumer data from the source data making data quality measurement at the consumer end unmanageable. The imperative for effective quality assurance at data source is becoming stronger as the consequences of poor data quality become more widespread.

Badri, Davis & Davis, 1995, pp2-4 claim that "*Quality management is a key factor in gaining competitive advantage*" when discussing quality measurement and management in a business environment with emphasis on "*top management leadership*". They also suggested that there was "…*no rationale was provided for the selections of factors included…*" and summarised noting that "*reliability and validity tests were very minimal*" and concluded that there are "...*8 critical factors (areas) of quality management in a business unit*", the seventh being '*quality data and reporting*".

Data quality management and associated issues are well known, documented and understood. The relationship between data quality and decision making is widely documented. The political, operational, social and financial implications of inappropriate data quality are well publicised. The operational risks presented by inappropriate data quality and the consequential decision making compromises are well understood in most environments especially and, more recently, in global financial environments.

Information quality issues are becoming more profound, more prevalent and better understood. As the cost of data storage continues to fall, the inverse relationship between cost and performance of processing power continues, and the breadth and depth of data volumes are increasing.

2.2 DATA QUALITY FRAMEWORKS

Data quality frameworks have been proposed based on theoretical, empirical and intuitive approaches (Price, Neiger & Shanks, 2008, p3) as the data quality issues have become better understood and the need the manage these issues have become more pressing.

Over the last 15 years data quality frameworks have progressed from ontological frameworks to context-based, systematically-based analysis frameworks.

Eppler & Wittig, 2000, pp84-86 analysed a series of information frameworks that have been developed between 1989 and 1999. Eppler & Wittig's approach has been adapted to illustrate the progressive theme that shows frameworks with a systematic problem solving approach. The notion that quality errors can vary for the same data when used in different contexts by different users becomes strongly evident from 2000 onwards.

Madnick, Wang & Dravis et al., 2003 and Madnick, Wang & Lee et al., 2009 describe an extension of the MIT Total Data Quality Management in the early 1990s that was based on research conducted by Madnick & Wang,1992, Juran & Godfery ,1999 and later Deming & Shewhart, 2008.

As data quality frameworks have been developed since the mid-1990s, recognition of users and their information usage has developed into a key theme that illustrates the difficulties associated with data quality management.

The differences highlighted in the list below are an acknowledgement that a common information component can present differing data quality levels across an organisation based on perceptions and usage.   This has become a fundamental development in data quality frameworks.

A Sequence of Data Quality Framework Development (based on a time series theme from Eppler & Wittig, 2000, pp84-86)

Table 2-1 A Sequence of Data Quality Framework Development

Author/s	Title	Data quality definitions or approaches
Wand & Wang, 1996	Anchoring Data Quality Dimensions in Ontological Foundations	Representation and interpretation where data presents a perceptible representation allowing inference about the real world.
Wang & Strong, 1996	A Conceptual Framework for Data Quality	Accuracy, Objectivity, Believability, Reputation, Accessibility, Security, Relevancy, Value-Added, Timeliness, Completeness, Amount , Interpretability, Concise and consistent Representation
Calero & Piattini (u.d.)	A data quality measurement information model based on ISO/IEC 15939.	Information Need, Measurable Attribute, Stakeholder usage, Measurement Methods, Random sampling (*of data*), Sampling and frequency of assessment
Alexander & Tate, 1999	Applying a Quality Framework to Web Environment	Authority validated information, author is visible, Accuracy, Objectivity presented without personal biases, Currency content up-to-date, clearly defined target audience and intuitive design.
Wang, 1999	A Product Perspective on Total Data Quality Management	TDQM Methodology to deliver high quality Information products just as any other manufactured product. Introduction of the notion that information quality needs to be assessed across roles, Wang p62, 1999
Knight & Burn, 2005	Developing a Framework for Assessing Information Quality on the World Wide Web	Notion that data quality as fitness for purpose in that data suitable for one purpose many not be suitable for another purpose. Wang & Strong,1996 and Tayi & Ballou,1998
Price & Shanks, 2005(a) Price & Shanks, 2005(b)	Empirical Refinement of a Semiotic Information Quality Framework	Analysis of criteria to assess an information quality framework using semiotic theory
Lei, Uren, & Motta, 2007, pp135-139	A Framework for Evaluating Semantic Metadata	Metadata evaluation approach based on rules that state: “Precise capture of data source meaning; Accurate representation of the real world; and conformity to underlying ontologies.”

Author/s	Title	Data quality definitions or approaches
		The SemVal architecture defines the need for an evaluation method, assessment metrics that define a quality model.
Burgess, 2007, Slide 6	Data quality description and examples	Notion that data quality is subjected to diminishing controls as it becomes more removed from the collection towards analysis and usage. Burgess also describes the thought that data quality may change its effective quality levels over time. Burgess also notes that at an individual level the same data can support many different tasks performed by different users. Data quality can then be perceived differently when used for different purposes and in different contexts.
Price et al., 2008	Developing a Measurement Instrument for Subjective Aspects of Information Quality	"... organizations must be able to monitor the quality of the information they produce or use, including both stored data sets and the information retrieved from those data sets. ...Essentially, the necessary foundation for IQ management is an effective means of defining and evaluating IQ."
Madnick et al., 2009	Overview and Framework for Data and Information Quality Research.	Describes *topics* and *methods* expanding the TDQM (Total Data Quality Management) program Madnick and Wang, 1992
Batini, Cinzia & Chiara et al., 2009	Methodologies for data quality assessment and improvement.	Data analysis, requirements analysis, identification of critical areas and measurement of quality. Critical area examines data flows and relevant databases.

Table 2-1 A Sequence of Data Quality Framework Development shows how perceptions and methods have changed to better reflect the issues surrounding the management of data quality.

The perceptions of data quality management have been changing and are expected to continue to do so as continued research is applied to the many issues associated with data quality.

2.3 The Changing Perceptions of Data Quality

DeLone & McLean, 1992, p67 argue that measuring IS (information systems) success is a combination of six components in an IS system (system quality, information quality, usage, user satisfaction, individual impact and organisational impact) and is represented as:

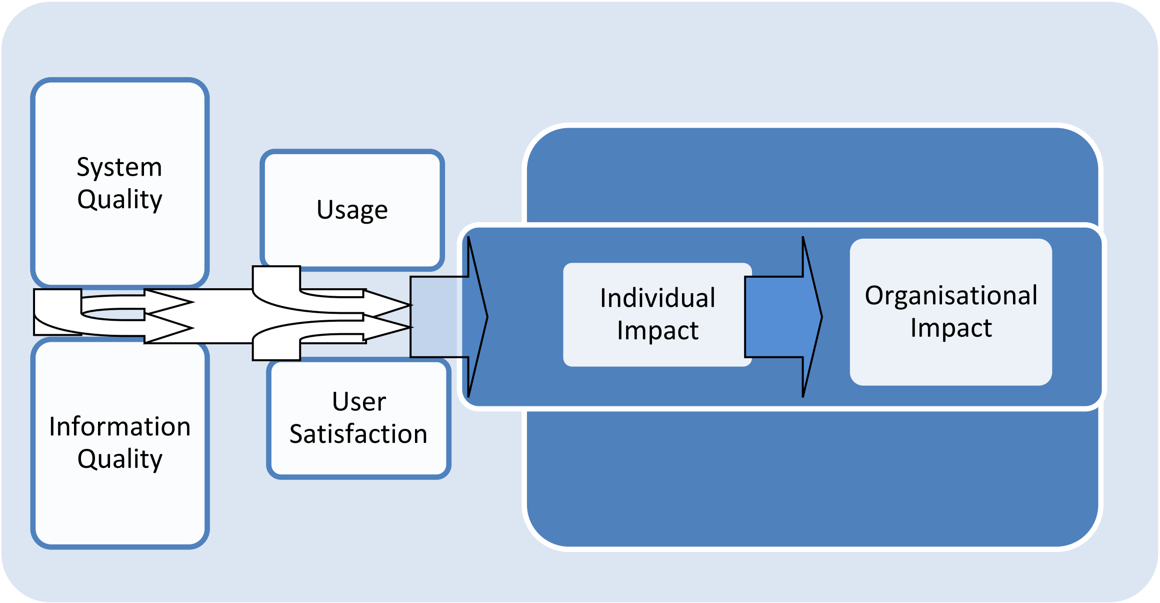


Figure 2-1 Information Quality and Organisational Impact
(*Adapted from DeLone & McLean,1992, p67*)

DeLone & McLean, 2003 describe using their *Model of Information Systems Success: A Ten-Year Update,* significant research that contributes to the information quality models. This research illustrates the thinking and the directions in which information systems and their components are subjected.

DeLone & McLean, 2003, p2 describe an evident direction using their model reproduced below. "*The role of IS has changed and progressed during the last decade.*" showing that information quality contributes or is directly causal to individual and organisational impacts.

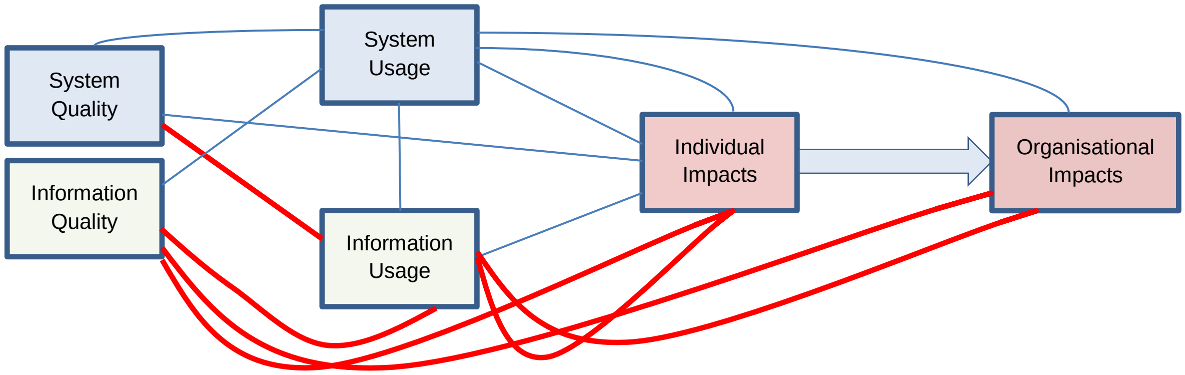


Figure 2-2 Dimension Association Relationships
(From DeLone & McLean, 2003, p14)

Figure 2-3 Dimension Association Relationships shows the relationship between information quality and impacts. The red link lines represent the space in which this thesis explores. This figure does not, however, include the ubiquitous nature of data quality as assessed by different users as they use information for decision making within an organisation.

The relationship between organisational impact, users, usage and information quality shows the information and system quality as mutually causal components.

The relationship between information and system quality, usage and user satisfaction, individual impact and organisational impact highlights the organisational imperative to manage the quality of information to an acceptable level.

Strong & Wang, 1997, pp136-137 add "*that quality of data cannot be assessed independently from the people who use (the) data*". The definition of data quality varies across varying frameworks and, as described by McKnight, 2005, p4 "*The currently accepted view of assessing IQ, involves understanding (it) from the users point of view*".

Pipino, Lee & Wang, 2002, p211 extend this theme stating "*Subjective data quality assessments reflect the needs and experiences of stakeholders: the collectors, custodians, and consumers of data products*" demonstrating that data quality issues are complex, user-relative and pervasive.

Information system owners are becoming increasingly accountable for information systems expenditure, information systems are becoming more comprehensive (lower cost of storage and supporting systems), and the implications of poor data are becoming better understood.

Tayi & Ballou, 1998, p54 note that "…*the use of legacy data in, for example, decision and executive support systems has refocused attention on information quality...*" They continue by describing the trend that data is now viewed as "*a key organizational resource and should be managed accordingly*". Tayi & Ballou do not suggest, however, that changing business models and directions can also render legacy data that might have been valid at one period, invalid in another. Legacy data can be flawed when viewed in the current data models. Information users may not be aware that legacy data has changed business relevance and may be presenting inaccuracies.

An information reusability theory is illustrated by Alstyne, 1999, p328 describes "*information goods*" by noting "*we cannot use information quantity as a direct input to either production functions or utility curves on the grounds that we cannot necessarily know when more is better*" showing that information (and its inherent quality levels) can be likened to 'reusable instructions' rather than tangible goods. Alstyne, 1999, p335 describes information goods using a value calculation showing that the value sum of sequence costs: $C = \sum_{k=0}^{n} c(a(i))$ (so the recurring opportunity to use information is inexhaustible and limited only by time) and that the limit is that of time (t) (to act upon new opportunities) that becomes zero $t \in [0,0]$ as $f(x)/x \rightarrow 0$ as $x \rightarrow 0$. So, then the opportunity (op) to use information is essentially limitless as t=> 0 then op $\rightarrow\infty$.

Even & Shankaranarayanan, 2007, p75 reinforce this perception stating that there is a "*need to revise data quality metrics and measurement techniques to incorporate and better reflect contextual assessment*". They note that "*Research has rarely examined data quality management from the economic perspective...*" (Even & Shankaranarayanan, 2007, p76). Here they highlight that there are financial imperatives and costs-at-risk as a result of data quality errors.

Even & Shankaranarayanan, 2007, pp75-93 do not, however, include the notion that there can be many contextual data quality measures for the same data at the same time.

When added to this theory the notion that data quality definitions can be subjective and variable, data quality can present many states simultaneously across an organisation. The shortfall in systematic, ontologically-based frameworks becomes apparent when used to describe the behaviour of information.

A common information component can have varying levels of quality measures and priorities depending upon the information user and their current information-related task. Much information can therefore have varying data quality levels when utilised by different users; or even the same users for different tasks.

This complexity adds to the required effort and management of data quality management in an organisation and can make effective data quality management a daunting task.

2.4 Data Quality in Supply Chains

In addition to the costs associated with data quality error detection and remediation within an organisation, there are further operational and economic risks associated with data quality errors external to an organisation in a supply chain operation, both internally and externally. The same issues that surround information within an organisation become more pronounced when viewed in a supply chain environment.

The Aberdeen Group, 2007, pp5-6 make a point in their white paper *"With so many manufacturers relying upon a global supply chain, visibility is necessary to identify and rectify quality gaps that can occur at any node in the supply chain, impacting both the cost of quality and finished product quality".* The Group continues, noting that in a supply chain, both the preceding and following links must have a common method for managing quality gaps with the information and production steps in the chain. The Aberdeen Group used this approach to develop the Best-in-Class PACE™ Framework (©Aberdeen Group, 2007, p10) by identifying Pressures, Actions, Capabilities and Enablers.

DataFlux Corporation, 2010, pp4-8 describe the effects of poor data quality in a supply chain "*…an organization's exposure to risk often leads … to a series of events that show the impact of poor data quality, such as an increase in customer churn, supply chain disruptions, or other events*" as a measure for an organisational maturity model. This measure has been developed by DataFlux who use these measures to determine an organisation's performance in a *Data Governance Maturity Model* using their Master Data Management enterprise view framework.

de Corbière, 2009, pp3-7 has expanded the notion of a supply chain (sequential model) using a PetriNet model described by Liu, Kumar & Aalst, 2007 showing the product supply, information sharing and relationships between organisations as four classes:

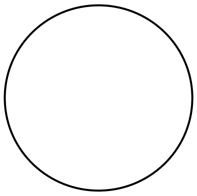
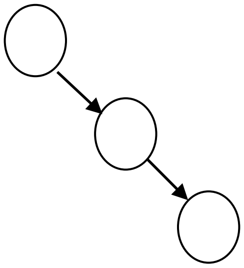
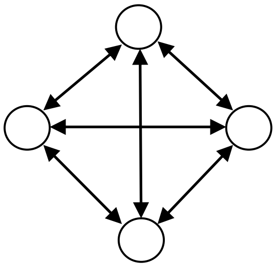
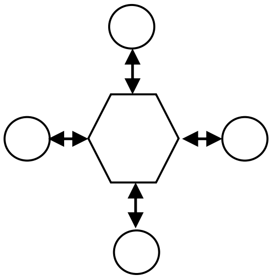
Interdependence	**Sequential**	**Reciprocal**	**Pooled**
			
Independent Organisational Structure (IOS)	Value/supply chain	Networked	Hub-and-spoke

Figure 2-4 Classes in the interdependence view

Liu et al., 2007, pp3-5 describes IOS (independent organisational structures) shown in *Figure 2-4 Classes in the interdependence view* that reflects data warehousing relationships within organisations.

This model shows various types of 'chains' all of which present data quality issues that affect either the chain or pool and reciprocal relationships. The Value/Supply chain is the classical supply chain model; the networked and hub-and-spoke model, although shown here to represent a reciprocal relationship between several organisations, can equally apply to data warehousing where multiple databases are related at some information level.

In this thesis, the type of relationship between organisations is not differentiated as the causal affects of poor data quality will still offer shared risks and costs through data quality shortfalls.

Indeed, organisations with data warehouses could also participate in supply chains offering multiple internal dependencies as well as external (and possibly multiple) dependencies.

Liu et al., 2007, pp764-769 uses Petri-net modelling to illustrate time-based Petri nets that illustrate IOS structures as event–based patterns. These patterns illustrate the event management associated with supply chains and offer the potential to illustrate the propagation and adverse event outcomes with data quality issues as the supply/receiver relationships (in any of the four models illustrated in *Figure 2-4 Classes in the interdependence view*) in order to operate.

Recognised and formalised supply chains are becoming more commonplace according to the Supply Chain Council (SCC), 2009 who have developed and enhanced the SCOR™ model first created in 2002, that reflects the "*extended enterprise*" and is used to model enterprise maturity models based on an organisation's management of quality workflows, manufacturing and information.

Data quality errors can be related to production, inventory, transportation and other operational measures as well as associated information products that relate to the goods or services that participate in the supply chain.

Detailed in The National Center for Manufacturing Sciences, 2002, the centre describes the SCOR™ model developed by the Supply Chain Council where they describe information flows that "*dictate daily operations*" showing that collaborative supply chains present an "*extended enterprise*". This model illustrates the multiplier effect of information quality problems both internally and externally.

This model was developed to reflect increased risks and costs associated with increased collaboration in supply chain and e-Manufacturing strategies.

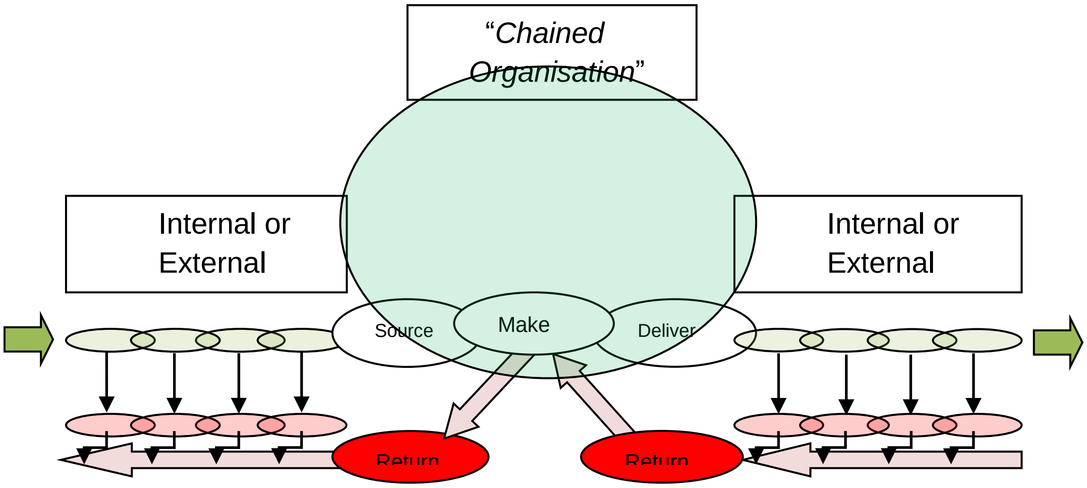


Figure 2-5 Extended Enterprise in a Supply Chain Relationship
(*adapted from the Supply Chain Council, 2002 SCOR™ MODEL*)

The model shown in *Figure 2-5 Extended Enterprise in a Supply Chain Relationship,* the multiplier effect of inappropriate data (intolerable data quality causing returns) in the supply chain using a "*Chained Organisation*" (*highlighted with a light green oval*) as an example both receiving and initiating returns. Examples here can include issues such as inappropriate requirements or specifications, unexpected delivery timing, inaccurate costings, unexpected quantities and so on. These information issues can present adverse outcomes for many links in a supply chain. An information-triggered return late in the chain could trigger multiple quality shortfalls and associated cost multipliers with the preceding chained organisations.

Notable in this model is that the later a data quality problem is detected in the chain, the greater the effect that this issue may exhibit against other members of the supply chain. Common examples of this cost multiplier are illustrated by the ACCC, (Australian Competition & Consumer Commission), 2011. Here examples of faulty component manufacture have manifested in end-product recalls that have cost the end-supplier substantial funds and resources.

A notable example in Australia is that of motor vehicle recalls from most of the major manufacturers who cite poor information exchange about components (ACCC, 2011). The costs associated with these recalls include not only rework, repairs or replacement but also the potential liability, lost reputation and the expenses associated with individual communication.

Standard and Poor's, 2004, pp4-5 have concluded in their studies that many organisations do not address the management of data quality detection and remediation in a systematic and cost effective manner.

While there has been a great deal of investment in information technology, the returns have not met expectations in terms of the quality of information offering a diminished return on IT investment.

Dasu et al., 2003, pp6-8 describe the changing views of data quality through knowledge engineering where data quality concepts are increasingly *"applied to databases that support business operations such as provisioning and billing..."* that assumes well understood business rules and their relationship to the supporting data. Business rules are often poorly documented and subject to (formal and informal) change with the need to gather business rules from subject matter experts, who may not always agree.

Redman, 1998, p80 notes that *"poor information quality increases operational cost because time and other resources are spent detecting and correcting errors."* Redman offers typical examples such as address correction; re-issuing bills and the time and costs associated with these '*rework*' activities.

The literature review conducted for this research shows this comment to be typical of a well-established understanding of the disadvantages associated with poor data quality.

Rudra & Yeo, 1999, pp1-6 when describing the emerging data quality issues with data warehousing, observed "*It has been found that often, many end-users, including managers are unaware of the quality of data they use in a data warehouse*". If there is no mechanism to inform information users, then their resultant decisions can be uninformed.

The problem Rudra & Yeo, 1999, pp1-7 describe is threefold:

1. Decisions may be compromised through poor quality information (believed to be valid); and

2. Managers may not be aware that their decisions may be based upon data with some level of quality errors; or

3. Managers may make decisions that do not reflect the information to hand as a result of low trust levels around valid data.

An emerging development is the increasing use of retrieved information ranking that relies on vector space models, probability models and fuzzy sets (Telang et al., 2007, p257). This emerging development places a greater reliance on appropriate data quality over an increasingly wide range of information.

Ross et al., 2007, pp2-5 describe how an organisation's (common) data sets can be differently ranked in importance by different users in different areas. This too highlights the need for data quality levels that meet varying needs across an organisation.

2.5 COST AND VALUE OF DATA QUALITY

Literature research shows that a significant number of organisations operate with poor quality information. Otto & Ebner, 2010, estimate that "*About 75% of organisations have identified costs originating from dirty data*". U.S. businesses pay $600 billion a year (estimated 5% US GDP) due to a lack of data quality (Eckerson, 2002, p5). O'Brien, 2011, p2 adds that (from Gartner

Inc. 2011 report) "*75% of organisations will experience significantly reduced revenue growth potential and increased costs due to the failure to introduce data quality assurance*". Clearly there is a pervasive and costly issue with unknown or unacceptable data quality.

Wang, 1996, p2 stated "*A recent industry report, for example, notes that more than 60 percent of the surveyed firms (500 medium-size corporations with annual sales of more than $20 million) have problems with data quality".* This issue does not appear to have been resolved in the last 15 years with the direct and indirect costs of data quality shortfall becoming better known rather than resolved.

Wagner & Meisinger, 2006, p38 describe the costs associated with discovery by '*breakdown*' using software quality assurance models using "*quality economics in general and in terms of the analytical model*". This model reflects the research by Strong et al., 1997, pp103-110 when describing the costs of attempting to detect and correct data quality errors.

'Current state' *Figure 2-6 Cost of Information - Current State* shows the perception that remedial detection and correction costs are no more cost effective than breakdown costs.

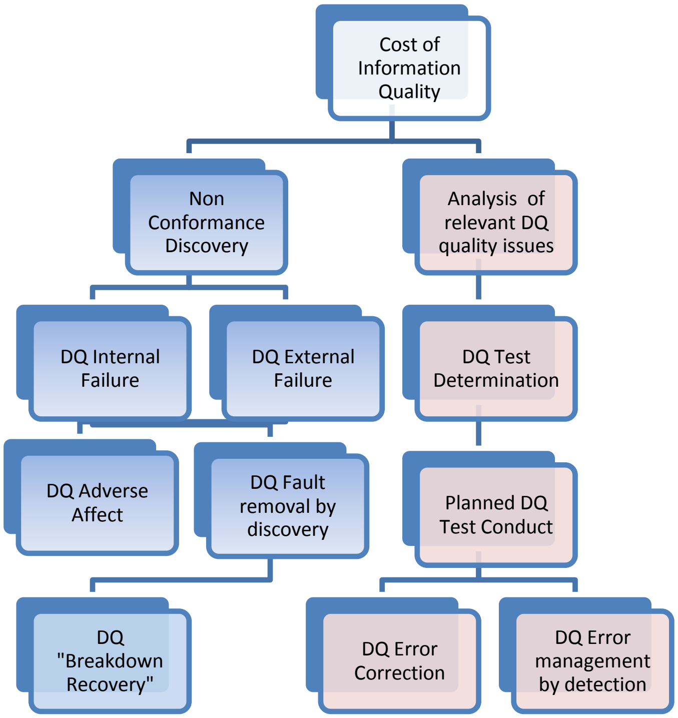

Cost of Information Quality
Non Conformance Discovery
Analysis of relevant DQ quality issues
DQ Internal Failure
DQ External Failure
DQ Test Determination
DQ Adverse Affect
DQ Fault removal by discovery
Planned DQ Test Conduct
DQ "Breakdown Recovery"
DQ Error Correction
DQ Error management by detection

Figure 2-6 Cost of Information - Current State
(adapted from Figure 1, Wagner & Meisinger, 2006)

This research proposes a demonstrable process that ranks data in order of technical and organisational impact. This process then allows data to be ranked, examined, tested and, if required, corrected based on each data's position in a relative data ranking. A percentage of critical data elements can then be subjected to testing and remediation based on their relative opportunity costs rather than selecting all data holdings as if they were all of equal importance to an organisation.

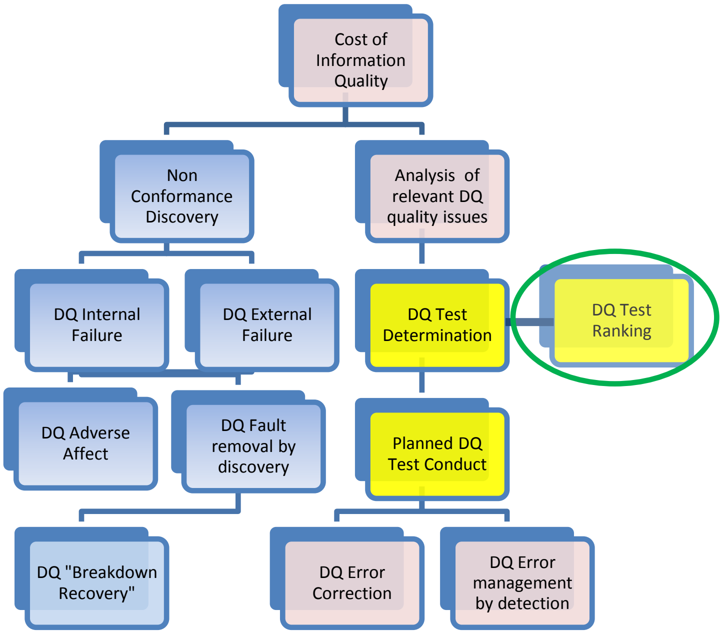


Figure 2-7 Cost of Information - Future State
(adapted from Figure 1, Wagner & Meisinger 2006)

The green circle in *Figure 2-7 Cost of Information - Future State* represents the proposed change to Wagner & Meisinger's 2006 model that qualifies the test determination approach. The DQ Test Determination and Planned DQ Test Conduct (DQ-Data Quality) are both effected to target investigation against highest ranked data.

Assuming that the impact of higher ranked data with quality errors offers higher failure costs than low ranked data, and then the scaling approach described in this research offers a level of advantage so that the "*breakdown recovery*" costs and "*adverse effect*" components could be reduced thus making strategic data quality measurement and management more attractive to organisations.

Alstyne,1999, p329 describes the determination of information value as "*notoriously difficult*". Alstyne,1999, p340 then describes the determination constraints as "*containment, abstraction, context, non-monotonicity, the inspection paradox, and the fact that it provides indirect rather than direct utility*".

Many authors have published research into the costs associated with poor quality information. The highlights of the change in thinking are illustrated below showing the awareness of the relationship to poor (and costly) decision making issues by many information users within an organisation.

Authors	**Theme**
Huff et al., (1995)	The value of information systems in society.
Alstyne, (1999)	*"The value of information to a decision maker as the difference between informed action and uninformed action. (and) as we change the decision problem, we change the context and thus the information's value.",* Alstyne,1999, p328
Xu, p628, (2000)	Managing the quality of accounting information has become critical. Identification of critical success factors in accounting information systems.
Delone & Mclean, (2003).	Evaluation of Information systems, development of "*Dimension Association tests*" model.
Stvilia, (2008)	"…*value-based assessment of metadata quality and construction of a baseline quality model*." Described using Dublin core metadata models as a "*representational object*". Stvilia describes the notion of "*the value of a quality change*".

Table 2-2 Information Costs

Kovac et al.,1997, p63 describe the link between data quality and an organisation's goals describing the corporate effort (as costs and resources) required to achieve an acceptable level of data quality. *"That quality of data is critical to organizations is a truism. Implementing a Total Data Quality Management (TDQM) program to achieve a state of high data quality, however, is not a trivial undertaking."* Given the business impact of data quality and the business expertise required to identify and plan for a data quality problem; the resources required to manage this process may be unavailable or perhaps even unwilling.

Iivari, 2002, pp8-20 describes how information systems are becoming pervasive in many aspects of human endeavors. Given the investments in IT and its ubiquity, the success of these investments is measured by the benefit to the organisation.

A survey commissioned by Pitney-Bowes, June 2009 interviewed 193 respondents of which 75 represented organisations with a turnover exceeding US$1 billion per annum. This survey shows some significant costs associated with data errors as well as 121 organisations who have not calculated the cost of errors at all (Waddington, 2009):

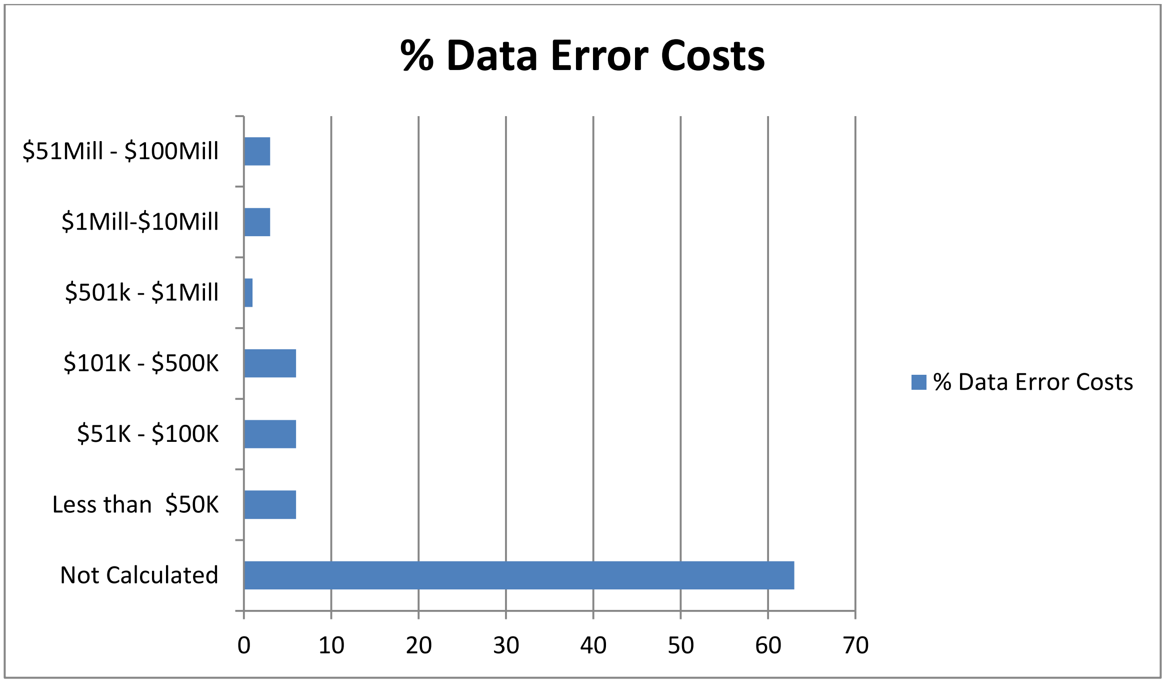


Figure 2-8 Survey of 193 Companies

(Adapted from a graph by Pitney-Bowes, June 2009)

Significantly, 60% of respondents reported that they do not calculate their costs at all. This outcome means that these respondents do not attempt to cost data quality errors nor do they check for data quality errors.

Waddington, 2009 also notes that 86/193 organisations do not attempt to measure their data quality; 60 at departmental level and 34/193 at organisational level. If the 9% (or about 17) who '*do not know*' are added to *'those who do not measure'*, then there is a significant percentage of the organisations surveyed who probably do not attempt to measure their data quality.

Waddington's comments suggest a corporate unwillingness to measure and then manage data quality issues. Given that the commercial impacts of poor data quality are well documented, why would so many organisations with a financial turnover exceeding $1 billion each not attempt to measure the magnitude of their problem?

Given that the estimate of 5% data quality problems appears typical (Standard and Poor's, 2004), (Cong et al., 2007, pp315-326), then this estimate suggests that at a financial turnover of $1 billion exposes each organisation to at least 5% x $1 billion at risk each year.

Gartner Inc. (from Moore, 2007, p1-2) when describing data quality issues estimated that "*three-quarters of large enterprises will make little to no progress towards improving data quality until 2010*" and that "*More than 25 Percent of Critical Data in the World's Top Companies is Flawed*"

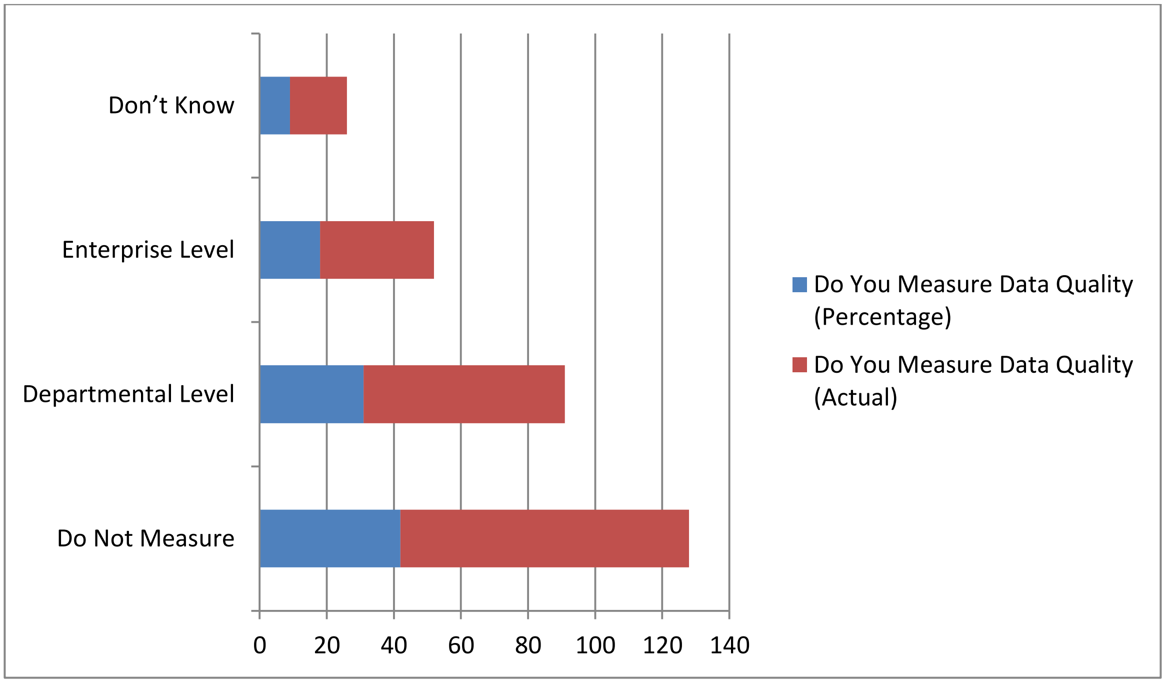


Figure 2-9 Data Quality Awareness in Organisations
(*Adapted from a graph by Pitney-Bowes June 2009*)

2.6 Data Quality Error Rates

Standard and Poors, 2006 estimated in 2004 that one quarter of the top 200 listed organisations in US reported *significant* data quality issues (*significant* = 5% or more) with their data holdings. S&P also estimated that the ROI (Return On Investment) generated around 14% returns in 2004 and approximately 33% returns in 2003. The 2004 returns were estimated to contribute 54% to the US economy.

Gartner, Inc., estimated that Fortune 1000 enterprises may lose more money in operational inefficiency due to data quality issues than they spend on data warehouse and CRM initiatives. Gartner Inc, (2007)

Powell, (2011) estimated that data quality problems cost U.S. businesses $600 billion each year and further notes from a survey across Europe and US that *"more than 61 percent of respondents admit they are currently making decisions based on half or less of their available data. Only four percent said they're using at least 75 percent of the available information".* This percentage when coupled with the data quality rates reported by Gartner Inc., (2007), offers a poor outcome for decision making.

Poor data quality in retail databases alone is estimated to cost US consumers $2.5 billion annually (Yakout et al., 2010, p2). The data quality costs associated with medical, administrative, and intelligence databases would be significant if the error rates are similar to the retail error rates noted above.

When describing the financial impact of data consistency and accuracy, Fan, W. et al., 2007, pp315-326 have noted that in a survey "*enterprises typically expect data error rates of approximately 1%–5%*”.

Can we assume that the remaining 75% of the top 200 organisations attempt to measure their data quality issues and do not have significant issues? This appears unlikely given the research offered by many authors and publicly-based institutions.

A recent audit report by the ACT Auditor General describes issues with the ACT Government's newly implemented Human Resources/Payroll facility saying that *"…system testing identified problems with unsatisfactory data..".* (ACT AG Report May, 2008 Section 1.6, p.8)
The ACT Auditor General also noted in the same report that *"…significant problems with data quality were experienced by most agencies."* (ACT AG Report May, 2008 Section 1.8, p10).
ACT Government analysts estimate that the cost of manual intervention costs to manage these data quality issues is costing approximately AU$1 million per annum in additional processing staff and an undisclosed amount in over-payment recovery and under-payment rectification.

Timmins, 2007, pp1-2 reported a critical statement about the UK Revenue and Customs department tabled by the (UK) National Audit Office (NAO) who "*detailed billions of pounds of overpayments in tax credits and another billion being lost to fraud and error. In its annual audit of the department, the NAO said overpayment in tax credits amounted to £6.6bn. Between £l.0 bn and £l.3 bn was paid in 2004-05 to claimants not entitled to it*". The NAO also reported that their estimates included *"£800m of tax due, while taxpayers are likely to have overpaid by £340m and potentially 5m people are not paying the right amount of tax".* These financial errors have been attributed to poor and variable information quality.

2.7 RANKING APPROACHES

There are no methods found that were found that rank information by value. There are, however, numerous ranking approaches described that reflect database and view ranking that offer some approaches to ranking data sets.

A successful approach to ranking related objects in an organisation is the Page Rank approach (Page, Brin, Motwani et al., (1998) and further enhanced by Langville, Amy & Meyer, (2006). A common commercial example of this accumulative ranking method is used by Google© who ranks web pages based on the number of links to the target page as illustrated below:

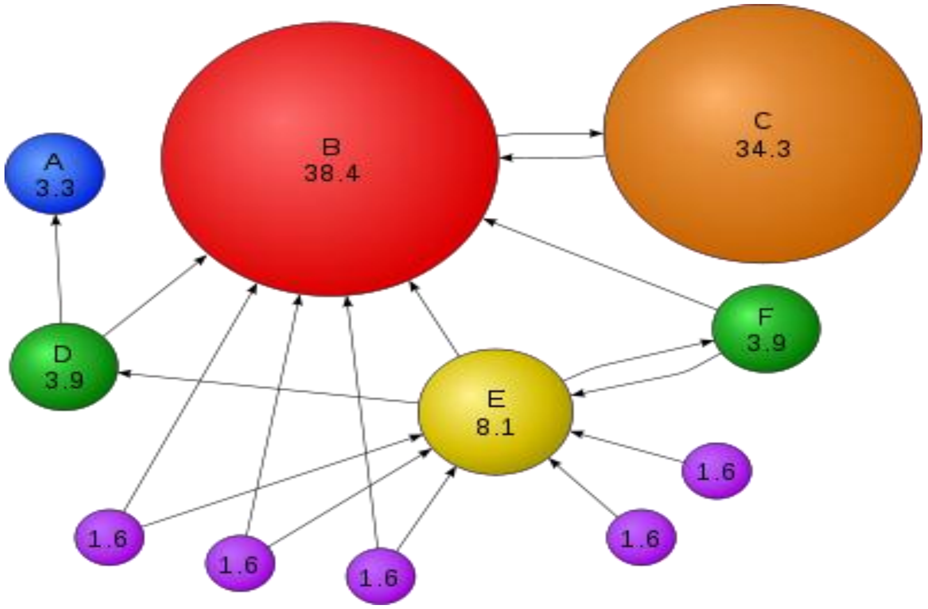


Figure 2-10 Google's implementation of Page Rank
(Page Rank application Google© 2001)

This approach offers an unlimited number of ranking depths in an hierarchical relationship relative to the target source page. Here a web page can be associated with many other web pages based on the hyperlinks present on each page. Data can be considered in a similar fashion where a particular data element in a table can be referenced many times in queries, reports and data entry forms as it is combined in various fashions with other data elements .

Haveliwala, 2003, pp784-789 proposes a method that improves on "*the single vector approach*". The '*topics*' are determined by the search expressions that Google uses to rank the page ('hits'). Haveliwala, 2003, p784 proposes that the ranking vectors be "*biased using a set of representative topics, to capture more accurately the notion of importance with respect to a particular topic".*

Could the Page Rank approach be applied to information valuation?

The research conducted by Zhou, Weston, Gretton et al., 2003, pp1-8 describes an extension to the PageRank ranking system with "*to rank data lying in the Euclidean space* ... *with respect to an intrinsic manifold structure*" that reflects a weighted ranking based on a '*network relationship*' that "*specifies the relative contributions to the ranking scores from neighbors*", (Zhou et al., 2003, p3).

Database ranking is an emerging trend that reflects the changing nature of information retrieval. The First International Workshop on Ranking in Databases describes *"In particular, a large number of emerging applications require exploratory querying on such databases;" as an alternative to "…Boolean retrieval modes"* (Ilyas & Das, 2007, pp49-50).   Ilyas & Das describe how databases might be ranked based on usage and TOP-K queries and then describe how database ranking approaches include *"dependency information in structured data…and keyword paradigms",* they conclude that a mechanism that offers *"Context-aware preference is introduced* *as a way to capture the changes in user needs and preferences with respect to the search context"* as an approach to *"complement traditional probabilistic information retrieval".* This work is in progress.

Might this approach offer an approach for valuing data elements based on dependency information within a database?.

"Data repair" (data quality correction) has been subjected to a ranking mechanism (Yakout et al., 2010, pp1-4) *where the approach has* been to rank data repairs based on "*quantifying the importance of satisfying a data quality rule, as well as, the benefit from a group of suggested repairs to the data quality*". This research does not, however, include organisational usage, value or prominence as determinants for Top-K ranking, but considers the *"value to the database"* in isolation to the organisation in which the database is used.

Bryant and Digney, 2007, pp1-10 describe *"Collecting and maintaining data are the duties of the system, rather than of its users".* This research includes a ranking value-based system, but considers the valuation to be technically driven. Whilst this approach offers a value-based indication, the addition of business usage and valuation is a key part of data quality management based on the notion that organisations value their data is decision making tool rather than the collection and storage of their data. The notion of organisational usage, value and prominence is not included.

2.7.1 *TECHNICAL WEIGHTING*

Ross, Stuckey & Marian, 2006, pp2-5 describe how an organisation's data sets can be ranked in importance differently by users in different areas; they do not include the technical propagation of data via views, reports and forms.

Heeren & Pitt, 2005 describe a ranking method as *"...historical ordinal data ranking as maximal claims..."* ('boasts') and offer a formula that optimises maximal 'claims'. This approach, however, assumes a single end-user of information rather than the potential for many users.

In addition, the outcome of the high ranked '*boasts*' does not include quality measures, but rather a best outcome based on a quantitative measure.

Feng et al., 2005, pp313-213 describe query ranking using multiple attributes based on dominance ranking. This ranking method describes a ranking mechanism that uses multiple criteria to derive an overall query ranking order to maximise large database searches. This approach might to be valid when scoring data items in a dataset.

Wagner & Meisinger, 2006, p39 propose an economic quality costing model by showing *"integration of a thorough stochastic model of the economics of analytical quality assurance..."* when describing a model for systems development.

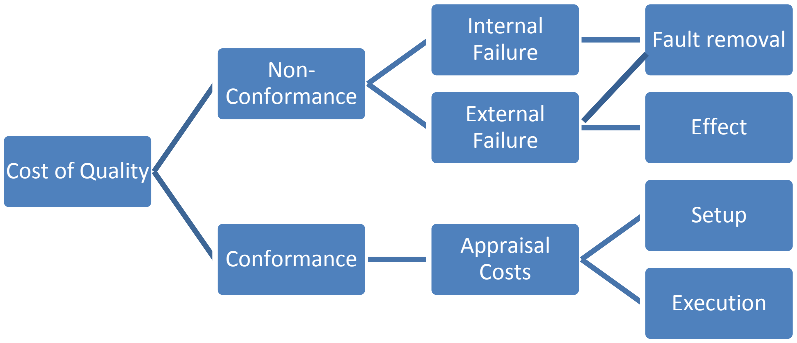


Figure 2-11 Costs of Quality

(adapted from Wagner & Meisinger 2006, p2)

Wagner & Meisinger 2006, p2 illustrate an analytical quality assurance model that essentially shows the cost of detecting errors, the cost of *not* detecting errors and the costs that are saved by detecting the errors.

These costs (and savings) are modelled by Wagner & Meisinger, 2006, p39 but assume that each defective data element is of equal value to the organisation.

Dunne et al., 2009, p851 describe an extension of Dung's argument theory where they describe an expansion against Dung's argument systems where '*attacks*' are weighted to indicate the strength of the attack. Dunne's framework describes an '*inconsistency budget that describes how much inconsistency we are prepared to tolerate".* Dunne et al., 2009, p853 then argues that that *"Weighted argument systems extend Dung-style abstract argument systems by adding numeric weights to every edge in the attack graph, intuitively corresponding to the strength of the attack, or equivalently, how reluctant we would be to disregard it.".* This argument supports the notion that defective data quality can exhibit various '*attack strengths*' (or harm to an organisation). Dunne's argument approach is an interpretation of a risk based discussion about 'damage' (risk impact), and 'attack strength' (risk probability). ' *reluctant we would be to disregard it'* expresses an organisation's appetite for risk.

In the data quality description context, the inconsistency budget describes an organisation's appetite for some level of quality defects. Dunne et al, 2009, p853 describe the notion of a weighted argument system stating "*we see that the use of explicitly numerical weights is under-developed*".

Lee et al., 2009, p1267, describe "*Query terms ranking is a research task aiming to rank a set of given query terms according to their effectiveness of retrieval*" as a ranking mechanism that offers to rank query terms in order of their effectiveness". Lee et al., then note that *"This ranking list is constructed by considering the effectiveness of a single term independently"* to demonstrate that various combinations of search terms can be used to rank queries with the higher the search ranked term, the higher the MAP score.

The MAP score is used to rank queries by their effectiveness based on key concepts (that would vary from search-to-search or user-to-user). Lee et al., 2009, p1268 finally note that "…*and can be well applied to other problems such as query expansion*".

This literature review shows that there are many and varied mechanisms for measuring the data quality in a system, but there is very little evident that describes approaches data ranking from a business and technical aspect. Individually, the ranking mechanisms described under Technical Weighting, p26 reflect different aspects that could, if combined, offer a repeatable and business focussed method for ranking data elements.

Although there are several ranking approaches described in Technical Weighting, p26 , there is no method for ranking all queries based on their contribution to decision making.

John Stuart Mill's method (Mill, 1846, pp479-658) ranks information causation in his logic, ratiocinative and inductive paper. This ranking mechanism classifies information collections into 5 methods ('canons'):

1.	Method of Agreement	-	Elimination
2.	Method of Difference	-	Elimination
3.	Joint Method of Agreement and Difference	-	Deductive
4.	Method of Concomitant Variations	-	Induction
5.	Method of Residuals	-	Induction

These canons classify information based on a causal (inductive) relationship between information.

In the context of this thesis, queries can provide information based on various levels of inference that is not immediately apparent. The degree of inference is the relationship and between atomic information sources (tables) and their relationships as defined in queries as the 'causal' relationship. Could queries in a relational database be ranked using Mill's the inductive relationship classification?

Classifying queries in the 5 information methods by adapting Mill's canons produces the following method.

Methods 1 and 2 (Agreement and Difference) well describe simple query filters. This has been further extended to describe simple queries that rely on one table that has an inclusion or exclusion filter.

Method 3 (both Agreement and Difference) describes a combination of filters and comprises one or 2 relational tables.

Method 4 (concomitant) describes a combination of filters and ordering or clustering and comprises 2 or 3 tables.

Method 5 (residuals) describes many tables (more than 3) and describes a combination of filters, sorted order and aggregation. Typically, the result query may produce few rows; or even no matching rows.

1. A query that comprises a single table and presents using a positive filter to produce a view;

2. A query that comprises a single table and presents using a negative filter to produce a view;

3. A query that comprises one or two tables and presents using positive and negative filters to produce a view;

4. A query that comprises two or three tables presents multiple filters and orders to produce a view; and

5. A query that comprises four or more tables, multiple filters and orders to produce a view.

2.7.2 ***BUSINESS WEIGHTING***

Abate et al., (1998) note that data quality is variable depending upon the data users and the user’s hierarchical position in an organisation “…*in terms of its conformance to its intended use”*. This comment discounts the value of automated *'one size fits all'* approaches as the importance of quality can vary between organisations and indeed within organisations for the same data sets.

Dunne et al., 2009, p854 proposes three weighting propositions.

Consider information as an argument and an attack is a description of some flaws (data quality errors) in that information :

1. **“*Weighted Majority Relations*:** *...one natural interpretation is that a weight represents the number of votes in support of the attack*”. Dunne et al., 2009, p853.

2. **“Weights as Beliefs: ...***Another interpretation would be to interpret weights as subjective beliefs*”. Dunne et al., 2009, p853. This interpretation describes the notion that a decision maker may believe information to be false when it is true or vice-versa.

3. **“Weights as Ranking:** Dunne et al., p853, 2009 notes “*A simple and obvious interpretation is to use weights to rank the relative strength of attacks between arguments... just the relative weight compared to the weights assigned to other attacks”.* Here Dunne is describing a scalar ranking approach, but does not illustrate how this weighting might be managed.

There appears to be a gap in the current thinking and research about the notion of valuing information and then using some determined value as a mechanism to rank and classify information in terms of its importance in an organisation. This ranked information can then be subjected to scrutiny as an approach to better manage data quality in an effective and efficient manner.

2.8 CURRENT SOLUTIONS

The current solutions appear to focus on the management of data quality issues as a reaction as they become evident.

Batini et al., 2009, p48 describes the Canadian Institute for Health Information (CIHI), 2005 "*to have been successful in meeting the primary objective of identifying and ranking the most critical issues of data quality improvement*". CIHI ranked their data holdings by assigning the most critical data quality errors at a higher ranking level than others. This method is claimed to offer some benefits when compared to a non-ranked approach. Batini et al., p50 (2009) state that a benefit has realised "*even though the number of evaluations is still low, numerous database improvements have already been implemented and many of these improvements might not have been detected otherwise*". This method is subjective and may vary with different user classes. This method states that some data quality errors are more critical than others based on the data *usage.*

There are many commercial packages developed to address the measurement of data centric issues in an information system by applying a set of rules against entire data holdings and comparing these measurements against consistency and format rules. See *Appendix B - Commercial Data Quality Providers* for some examples of major commercial package providers.

These commercial packages offer to address quality issues across the entire set of data holdings as potential quality assurance targets. These packages use sophisticated automated and technical measurement solutions to a perceived technical problem. Data may be measured to present an acceptable quality when evaluated against consistency or accuracy across some column definition or 'within boundary', but can still present poor quality when compared to the organisation's data usage by different users.

Batini et al., (2009) describe the distinction between *data-centric* and *process-centric* quality errors when classifying data quality measurement strategies as *data driven* and *process driven*.

Data-driven strategies focus on data values. For example, out-of-boundary data can be updated by refreshing against a more current database; postal address information can be verified against a postal service database; name-based data can be checked against electoral rolls and so on.

Data-driven approaches lend themselves to automated detection and possible consistency remediation.

Process-driven strategies, however, address the processes that create, modify or combine data to present information sets. For example, processes can be analysed and redesigned to emphasis activities that controls the source, access and format of data before and after storage. These activities tend not to lend themselves to automation as the processes and management can vary across an organisation for the same data sets.

Batini et al., (2009) cite key data quality measures "*six most important classifications of quality dimensions*". These dimensions have been noted in table 2-3 showing the differentiation between process and data driven data quality measures.

These six classifications have been expanded to show which classifications are data drive and process driven.

Table 2-3 Six Most Important Classifications of Quality Dimensions
(Modified from Batini et al., 2009)

Dimension	**Test Type**	**Data Driven**	**Process Driven**	**Examples**
1. Accuracy	Comparison	X	X	Can the data be verified against a discrete data source? How? Is the data *sufficiently* accurate to suit any classes of user in a particular organisation? Are there different levels of accuracy across an organisation?
2. Completeness	Inspection	X	X	Are there missing components in some collection of data. Are these missing components important to any classes of user in a particular organisation? Are there different classes of inspection?
3. Consistency	Inspection	X		If like data sets exhibit differences are they expected or outliers? Who would know and how important is this knowledge?
4. Timeliness			X	Often a tension between accuracy, consistency and cost. Are the timeliness needs consistent across an organisation?
5. Cost Effectiveness			X	Is the expenditure appropriate to the value and risks of the decision making that is supported by some data class. Given that most organisations

Dimension	Test Type	Data Driven	Process Driven	Examples
				have limited resources; are they being deployed to offer the best outcomes?
6. Availability			X	This can be a technical or logistical issue. Indeed the concept of availability can vary between different stakeholders at different times.

Process driven data quality errors are more likely to exhibit variable errors as data may be free from logical errors and still exhibit process driven errors; perhaps for some users and not others.

The key is that *process driven* data error detection offers significant and variable difficulty when compared to *data driven* quality issues.

The major cost of data quality management is the detection and measurement of data quality issues. Standard and Poors (2004) note that some large organisations are aware of their data quality problems and the associated opportunity costs associated with poor data quality.

Welzer, Brumen, Golob et al., 2002, p2 when describing data quality (in a medical context) "*Data quality has syntactic and semantic component; the syntactic component is relatively easy to achieve if supported by tools (either off-the-shelf or our own), while semantic component requires more research*" . Welzer et al. continue noting that "*Either semantic or syntactic incorrectness can have fatal consequences*". Welzner et al. also describe semantic data quality as the *properties* of data quality and then the syntactic quality as the *meaning of the data* which could be from legacy systems where the system has embedded data rules or the business has changed significantly so that the original intended (or understood) meaning has changed. Welzner et al., 2002, pp1-4 note that the semantics of data quality are well understood, but the syntactic quality and usage require more investigation.

(It is noted that 'syntactic' and 'semantic' are reversed in Welzer's et al. paper, perhaps as a typographic error).

Wang, 1999, p57, notes in his foreword "*To increase productivity, organizations must manage information as they manage products*". This comment states the organisational value of information, but not how it might be best managed. Comparing information to products highlights its importance, but does not account for the reusability of information.

Combining the notion of infinite reusability of information with comments from Knight et al., 2005 and Strong & Wang, 1997 (who illustrate that the data can present different quality perceptions from user to user) simply assessing and addressing data quality as a single attribute for an organisation cannot be effective.

Even, 2007, pp75-93 expands this concept by showing a value-based relationship between an organisation's data base collection where he shows that data usages with a "*high utility"* will have a "*more significant effect on the embedded value of a record"* as a function of usage context. Even then concludes "*Data consumers assess quality within specific business contexts or decision tasks. The same data resource may have an acceptable level of quality for some contexts but this quality may be unacceptable for other contexts".*

A practical illustration is shown by The National Centre for Manufacturing Services (TNCMS) and University of Michigan, 2002, pp1-26 when describing the cost of information assurance, propose an asset-based approach as organisations increasingly share collaborative information. TNCMS, 2002, p6 note that "*Despite intent to increase collaboration, manufacturers have often overlooked one of its most critical components, Information Assurance*". This comment describes the "*limitless reusability"* of information between commercially related organisations.

Hasan & Padman, 2006, pp324-326 explore the option to simulate data sets from various sources as a clinical decision support system. Current clinical decision support systems present quality reliability (accuracy and completeness) as low "*one study shows that accuracy and completeness in medical registries may be as low as 67% and 30.7%, respectively*". Hasan & Padman, 2006, p324) also describe the outcome of this study as presenting the risk of "*negative patient outcomes*". Hasan & Padman continue by suggesting the development of a data simulator "*To achieve this aim, we employ a two-pronged approach, first using a simulation model of data generation, data adulteration and CDSS use, followed by regression to quantify the impact of each data element on overall CDSS accuracy*."

Conclusion

The literature review chapter has highlighted multiple issues with the management of data quality, both in terms of opportunity costs, management costs, considerable complexity and effort.

Although the issues are well documented, many large organisations (turnover in excess of US$1 billion) agree that they do not address data quality and note that it costs them considerably in terms of rework, lost sales, lost product and considerable loss of goodwill.

A major constraint that contributes to the difficulties associated with data quality error detection is that *process driven* data quality issues that vary from user to user as well as when used to inform different decisions. For example, the same information may exhibit different process errors to different users at the same time.

The underlying issue is then that the cost of determining data and then detecting these errors is a key constraint to effective data quality management.

There is considerable research into various aspects of data quality weighting and ranking with piecemeal solutions suggested.

A comprehensive solution that calculates data weighting based on processes, different classes of users and usage has not been found.

Although data quality simulation has been proposed there does not appear to be any working models that offer this facility.

3 THE RESEARCH PROBLEM

Chapter Introduction

This chapter describes the problems, motivation and the gap that has been addressed by this thesis.

3.1 THE PROBLEM IDENTIFIED

The literature review has canvassed the primary dimensions of the data quality domain to be:

1. Data Quality and Decision making (See *Data Quality and Decision making*, p7);
2. Data Quality Frameworks (See *Data Quality Frameworks*, p9);
3. The Changing Perceptions of Data Quality (See *The Changing Perceptions of Data Quality*, p12);
4. Data Quality in Supply Chains (See *Data Quality in Supply Chains*, p14); and
5. Cost and Value of Data Quality (See *Cost and Value of Data Quality*, p18).

This review reveals a range of issues that need further research but an outstanding one, the one addressed in the research described in this thesis, is to determine which data have the highest impact in an organisation thereby allowing data quality assurance to be better targeted.

If automated solutions such as those referenced in *Appendix B - Commercial Data Quality Providers* can offer an effective data quality solution, then why does data quality remain a problem in a substantial proportion of the top 200 US corporations? (See *Cost and Value of Data Quality*, p18).

Automated solutions can detect data driven problems as these types of problems can be detected by comparison with other databases; rules based value management and other consistency detection.

Much more problematic are process driven data quality errors. These types of errors are more difficult to detect using automated detection facilities as they data may match various consistencies and logical rules making it undetectable using automated means.

If the costs associated with the detection of various data quality problems represent a significant portion of correction costs, is this the reason for avoiding the data quality analysis or reacting to crisis-driven issues only?

A poor detection rate when testing for data quality issues may present an apparent poor return on investment.

For example, using a sample database of 20000 contestable data elements, a 5% error rate means that the gap between ‘strikes’ (true positives - ie. data element with a data quality error) could be significant with twenty (20000/1000) data elements examined between ‘strikes’.

The costs associated with examining a data element to determine the level of data quality error would logically be the same for data regardless of the detected quality state. Therefore the sunk costs associated with data quality error detection can be high with poor apparent return. Using this example, the cost of detection would be nineteen 'false' outcomes for just one ‘positive’ outcome. Should the data element detected is of low value to an organisation and with the subsequent error remediation effort, the effort and costs associated with this detection may not appear worthwhile.

Can data be ranked by assigning some numeric to data elements to reflect organisational value in order to address these examination costs?

3.2 MOTIVATION

The problems associated with inappropriate information quality are extensively documented. The problem is that the effort required to manage information quality against all data holdings does not appear to offer appropriate benefits. (See 2.5 - *Cost and Value of Data Quality*, p18)

A consequence of this constraint is that data with quality issues may be subjected to any of the following remedies:

- Not selected at all;
- Selected for quality assurance analysis based on historical understanding;
- Selected as data driven errors using an automated facility; or
- As disaster-driven reaction behavior.

The financial and governance imperatives for organisations to demonstrate an acceptable data quality level are clear and well documented. There are many methods for addressing data quality; so the question becomes *“why is data quality either poorly or not managed?*”

Literature research shows that the issues associated with poor data quality are well understood in public and private organisations.

Perhaps the time and labor required to manage data quality is a combination of resource availability, deployment expenditure and resource costs. Is this because the preventative costs show that the risk cost of data quality outcomes compared with costs associated with detection and remediation may appear to be viable. (See 2.5, *Cost and Value of Data Quality*, p18)

Many organisations appear to have adverse levels of data quality that, in turn, offer an adverse effect on their operating costs and profits or public performance; data quality issues remain pervasive and common. (See 2.5, *Cost and Value of Data Quality*, p18)
The commercial and governance benefits of acceptable (or at least known) data quality are well documented. (See 2.1, *Data Quality*, p7)

There are many commercial technical solutions noted in *Appendix B - Commercial Data Quality Providers* that address a variety of data driven quality problems, but these automated solutions are geared to consistency and referencing against other databases. These solutions test and correct mechanically detected data-driven data quality issues.

3.3 The Knowledge Gap

The problem appears to be that the effort expended measuring and managing data quality against all data holdings may not offer a fair return on investment (ROI). There does not appear to be a domain sensitive ranking mechanism that allows data to be prioritised for quality analysis addressing both data-driven *and* process-driven quality problems.

Delone & Mclean, 2003, p14 present a model of Dimension Association tests shown in *Figure 3-1 Dimension Association Tests*. The area overlaid with green ovals shows the gap this thesis is addressing as the problem and the motivation. The green connector lines show the relationship between the dimensions.

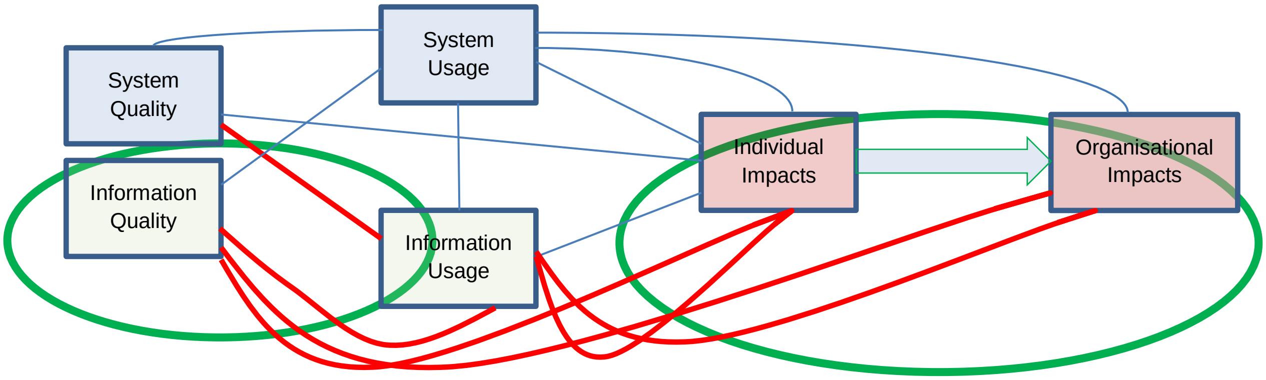


Figure 3-1 Dimension Association Tests

Adapted from Delone & Mclean, 2003, p14

3.4 The Research Question

To address the findings from the literature research, the research question becomes:

"Is there a method that allows an organisation to identify data that, should it have data quality issues, presents the greatest risk to an organisation?"

A key solution to this question is then *"Can an artifact be designed and developed to allow data to be assigned a value using a valuation system?"*

3.5 Chapter Conclusion

This thesis describes a method that allows high-value data to be identified by ranking database data and then testing the effort return by sorting all database data in value order thus allowing the highest-ranked data elements to be assessed.

The outcome of this approach is expected to show that rather than focusing on numbers of data quality errors detected (and possibly correction), the focus becomes detection of data quality errors that, if positive, present a high damage or threat impact to an organisation. The measure of effectiveness then becomes the cumulative scale ‘damage’ values of data errors detected and addressed rather than just data quality error counts.

Different error rates can be simulated to allow experimentation against different scaling approaches whilst allowing detection and correction rates to be tested.

4 Research Design

Chapter Introduction

This chapter describes the research method used to support the iterative design of a data ranking system that is to be used to conduct the experiments against the research question (See *The Research Question*, p38).

This research design uses a set of design science processes that reflects a cyclic and iterative artifact development and testing sequence. This set offers a framework for the development of software that proves the concept and then expands the ranking concept to provide multiple measurement points as well as simulation of data quality errors in a target database.

4.1 Design Science Process

The design science theory proposed by Peffers, 2005 and then expanded by Peffers, Tuunanen, Gengler et al., 2006, pp89-91 describes the relevance to information systems (based on other disciplines) of design science elements, namely Concept, Outcome, Design & development, Demonstration and Evaluation. This theory, whilst describing the overall concept, does not offer a framework for iterative software development.

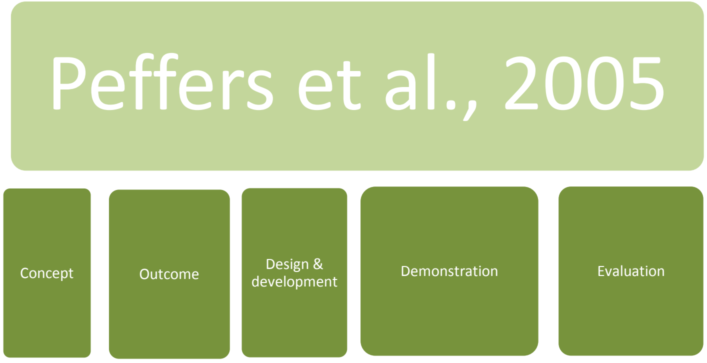


Figure 4-1 Peffers et al., 2006 Design Science Framework

Hevner et al., 2004, p3 describe a design science method "*using an IT artifact, implemented in an organizational context …perceived usefulness, and impact on individuals and organizations (net benefits) depending on system, service, and information quality*". Hevner et al.,2006 describe five evaluation methods: "*observational, analytical, experimental, testing, and descriptive*". This approach is used to describe the artifact development component of the approach.

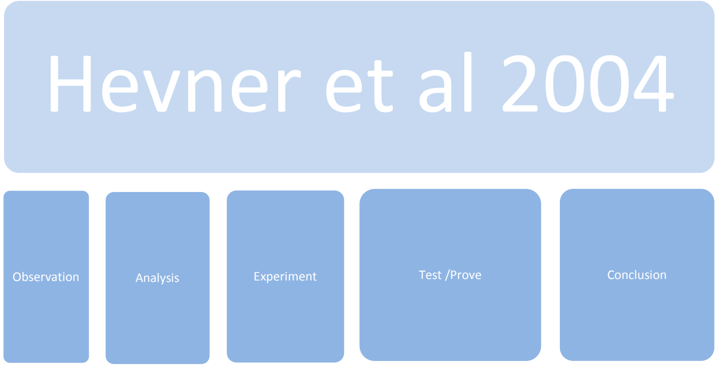


Figure 4-2 Hevner et al., 2004, DSR sequence for an artifact

This model describes the design and development component that was then adopted to create the Ranking Tool artefact.

The two models are then combined showing the framework from Peffers et al., 2005 with the software artefact design from Hevner et al., 2004.

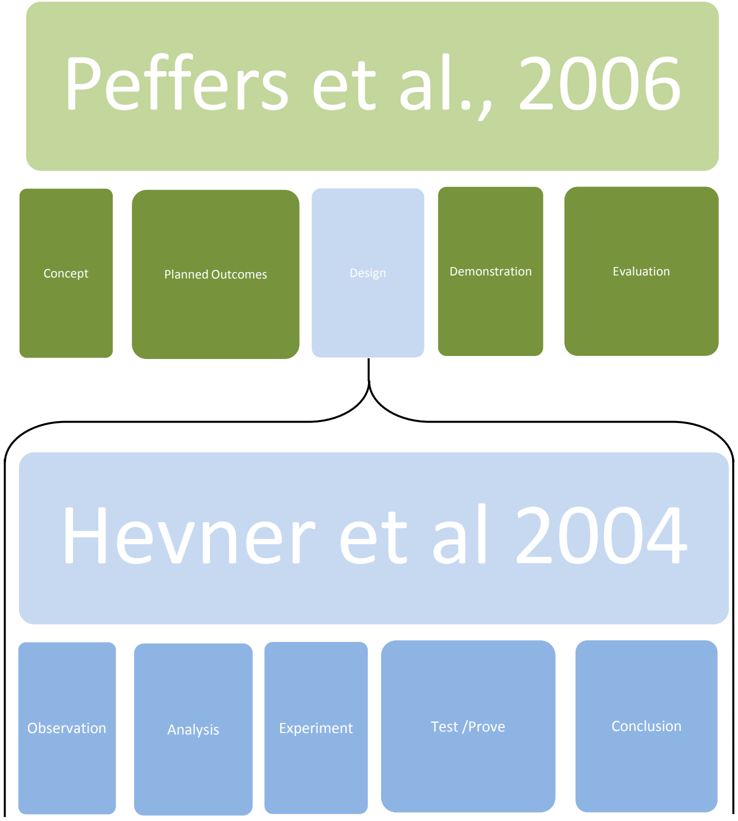


Figure 4-3 Peffers, Hevner DSR Combination

Given that the approach, the artefact and the subsequent experiments examine the notion of quality assurance both against the notion of object quality (the software artifact) and subject quality (the outcomes), the objects and subjects generated and produced using this design science process an overarching quality assurance method such as described by Bartneck 2009, p5 who posits that "*Quality is in the objects and subjects at the same time".* This definition describes the nature of information (and its quality measures) where information is both an intangible object and a target subject at the same time as well as relative to the current user who is using the information.

Baskerville, Pries-heje, & Venable (2008) emphasise the importance of evaluation, both before the design and development of an artefact as well as after the development of the artefact. Baskerville et al., 2009, p2 later note that design science is evolving into the "*evaluation of design science outputs, including theory and artefacts*".

The final design science model is represented in *Figure 4-4 Peffers, Hevner & Bartnek DSR combination*:

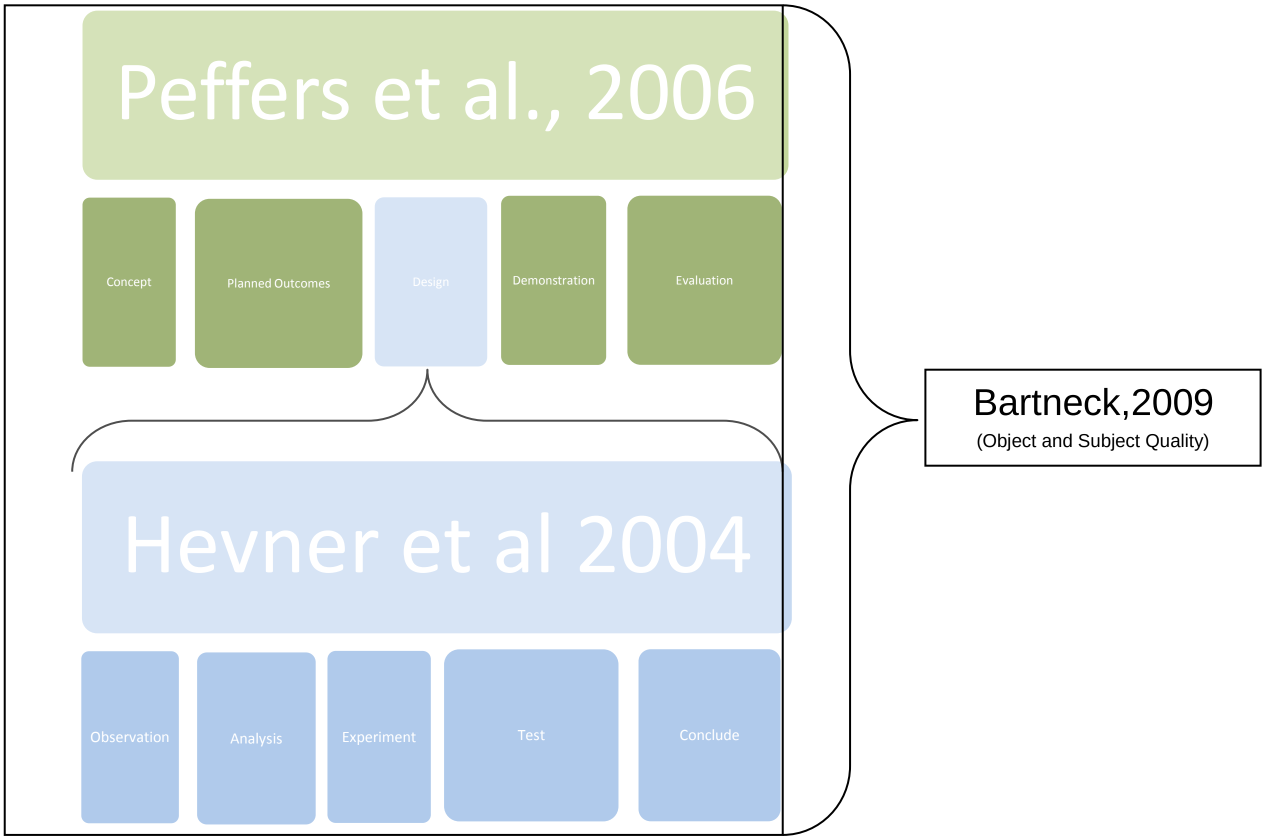


Figure 4-4 Peffers, Hevner & Bartnek DSR combination

In this thesis, the development of the theory and artifact design was iterative with pre and post evaluation in a reiterative series of cycles.

Offermann, Levina, Schönherr et al., 2009, p3 describe the principles of an '*iterative micro process of learning and designing*". These steps describe the development and testing process that the artefact underwent to support this thesis.

To illustrate the use of these models, the design science research (DSR) values used in this thesis are highlighted in red in *Figure 4-1 Peffers et al., 2006 Design Science Framework* .

Variable	**Value**				
Approach	Qualitative	**Quantitative**			
Artifact Focus	**Technical**	**Organisational**	Strategic		
Artifact Type	**Construct**	**Model**	Method	Instantiation	Theory
Epistemology	**Positivist**	Interpretivist			
Function	Knowledge	Control	**Development**	Legitimization	
Method	Action research	Case Study	**Field Experiment**	Formal Proofs	
	Controlled Experiment	**Prototype**	Survey		
Object	**Artifact**	Artifact Construction			
Ontology	**Realism**	Nominalism			
Perspective	**Economic**	Deployment	**Engineering**	Epistemological	
Position	Externally	**Internally**			
Reference Point	**Artifact against Research gap**	Artifact against the real world	Research gap against the real world		
Time	Ex Ante	**Ex Post**			

Table 4-1 Variables and Values for DR Artifacts
(From Cleven et al., 2009, p3)

The concept proposed is that rather than simply examining a database for data quality errors and then correcting them, this approach informs the analysis, design and development, trials and finally experiments against a ranking mechanism

4.2 The Concept

Much literature describes (See,*The Changing Perceptions of Data Quality*, p12) the issues that organisations experience with measuring data quality issues. The key issues are the effort and costs associated with testing for data quality errors for a poor apparent return. (See 2.5,*Cost and Value of Data Quality*, p18).

The concept is to design an artifact that ranks data in a database (using both technical and business measures) and allows the highest ranked data to be displayed and subjected to data quality evaluation. The aim is to demonstrate the value of testing data elements for data quality; both unranked and ranked, to determine if there is an advantage to examining and managing high-value data rather than unknown-value data.

A public domain database that represents organisational reports, forms and hierarchical structure was chosen, Microsoft’s North Wind Trader’s database.

The aim is to score each contestable data element at various measurement points as data is entered on entry forms, manipulated through views, and used in reports. Once all contestable data has been traced and scored, it is sorted to show high-ranking data first in a list of data.

The target database is “seeded” to simulate various data quality error percentages and the outcomes observed using different states and the effectiveness of each state measured.

Varying percentages of data errors are ‘cleared’ and the effect on the database observed to test the difference between states 1, 2 and 3:

1. **Current State** - This state is considered to be the current process for detecting data errors. The ranking values are accumulated as data quality errors are detected and cleared. The effectiveness of this approach is measured by accumulating the score value of data elements with errors and the number of data elements detected with quality errors and comparing the same detection process with the Ranked State;

2. **Ranked State** - This state is proposed as the a more effective method of managing and detecting data quality errors;

3. **Perfect State** - This state is treated as a control. The data is ranked in order of scalar value (highest to lowest) and only for data elements that have data errors.

Given that there appears to be no commercially available artifact that offers this ranking function, a software package has been designed, planned, developed and tested to offer an experimental platform that allows the experiments noted above to be conducted. (*Appendix F - Attached CD*) for details about the developed software system that presents the design science research artifact.

4.3 DEVELOPMENT HISTORY

A successful approach to ranking similar and related objects in an organisation is described in 2.7, *Ranking Approaches*, p24

This ranking method offers an unlimited number of ranking depths in an hierarchical relationship relative to the target source page. Here a web page can be associated with many other web pages based on the hyperlinks present on each page. Data can be considered in similar fashion where a particular data element in a table and can be referenced many times in queries, reports and data entry forms as it is combined in various fashions with other data elements .

This Page Rank© method was modified so that each contestable data element can acquire a value based on ranking points at different levels using the single vector approach. The accumulated ranking can then be ‘attached’ to each contestable data element thus allowing an ‘*in toto’* ranking comparison of all contestable data elements as a result of this relationship.

Much of this initial analysis has been focused on the work by Dunne et al., 2009, p854 who discussed the notion of weighting representing the “..t*he number of 'votes' in support of the 'attack'*”. (See 2.7.2, ”Weighting", p29)

This method allows data to accumulate 'votes' thus presenting each data elements' 'attack' value. The higher the accumulated value ('votes') of a data element, then the higher the value of that data element in an organisation and then the more vulnerable it would be to a quality deficiency (or 'attack').

4.4 Solution Objectives

The outcome of this research is to design and demonstrate a priority mechanism that allows data to be ranked in such a way that core data is exposed and subjected to data quality analysis and some appropriate remedial action. Core data is data that is considered essential to an organisation.

This thesis describes the development of a numerical weighting mechanism that incorporates all three weighting mechanisms (Dunne et al., 2009, p854) in order to rank data elements:

1. **"*Weighted Majority Relations*:** Dunne et al., 2009, p853 where Dunne describes the relative weights of 'attacks'.
2. **"Weights as Beliefs: ...***Another interpretation would be to interpret weights as subjective beliefs*". Dunne et al., 2009, p853.
3. **"Weights as Ranking:** Dunne et al., 2009, p853 notes "*A simple and obvious interpretation is to use weights to rank the relative strength of attacks between arguments... just the relative weight compared to the weights assigned to other attacks".*

These three weighting mechanisms have been incorporated into the ranking artefact by including rank points from both a technical as well as varied usage aspects.

This thesis proposes that instead of raw data quality defect counts, the *ranked value* of contestable data elements could be used as the metric for reducing high impact data quality errors to an acceptable level. This approach then modifies Wagner & Meisinger's 2006, p38 formulae to reflect the notion that data quality varies depending upon the user and their data usage.

The ranking should then, encompass both the technical relationship of data sets as well as the business relation of the corresponding information. This approach matches Dunne's weighting arguments using Wagner & Meisinger's, 2006 economic approach for coding defects in a system.

The solution is intended to improve effectiveness associated with data quality measurement and management. This research examines the target database using a developed artifact:

- a means to identify a database's propensity (based on structure and usage) for data quality error propagation;
- a means to identify and test a ranking method;
- a means to establish the significance of different ranking points as the result of a sensitivity analysis;
- a means to identify high-value (high ranked in descending scalar order) data thus allowing targeted examination and correction; and

- a means to identify low-value data that may not be cost effective to collect, store and archive.

Chapter Conclusion

This chapter shows how a modified design science method has presented a framework for the artifact design. The experiments have been designed in response to the gaps identified in the literature research.

5 SOFTWARE RANKING TOOL DESIGN AND TESTING

Chapter Introduction

This chapter describes the processes that were undertaken to design and test the supporting ranking tool software.

The process shows five of the major iterative steps that were undertaken to develop the current artifact. There were minor iterations (design, write, test) throughout the development that reflected conventional software development and code testing.

Each iteration follows the pattern outlined by Hevner et al., 2004, p3 who describes a design science approach "*using an IT artifact, implemented in an organizational context …perceived usefulness and impact on individuals and organizations (net benefits) depending on system, service, and information quality*". This approach offered continuous enhancement to better address the research questions.

Each iteration is shown as a cycle that informs the next iteration. The evaluation components within each cycle are: observe, analyse, experiment, test and conclude.

Iteration 1 – Proof of Concept

This iteration describes the initial trials to determine that a database can be automatically analysed and additional properties for each contestable data element could be defined. Once this proof-of-concept had been completed and tested, it then informed a trial ranking facility.

Iteration 2 – Initial Modification

Initial modifications were then designed and implemented to show that four primary ranking points could be evaluated. Once these ranking points were tested, the ranking system was ready for ranking point calibration.

Iteration 3 – Ranking Point Calibration

The modifications were then subjected to calibration tests. Preliminary analysis was conducted to test the viability of the enhancements using statistical measures. These tests showed that the ranking software was viable and supported the ranking value accumulation for data elements. The ranking system was ready for further ranking points to be added and the system recalibrated.

Iteration 4 – Six Ranking Points added and recalibrated
The analysis and testing conducted in Iteration 3 was used to inform the addition of two new ranking points with additional recalibration and statistical testing conducted. The ranking system was now ready for experimental trials.

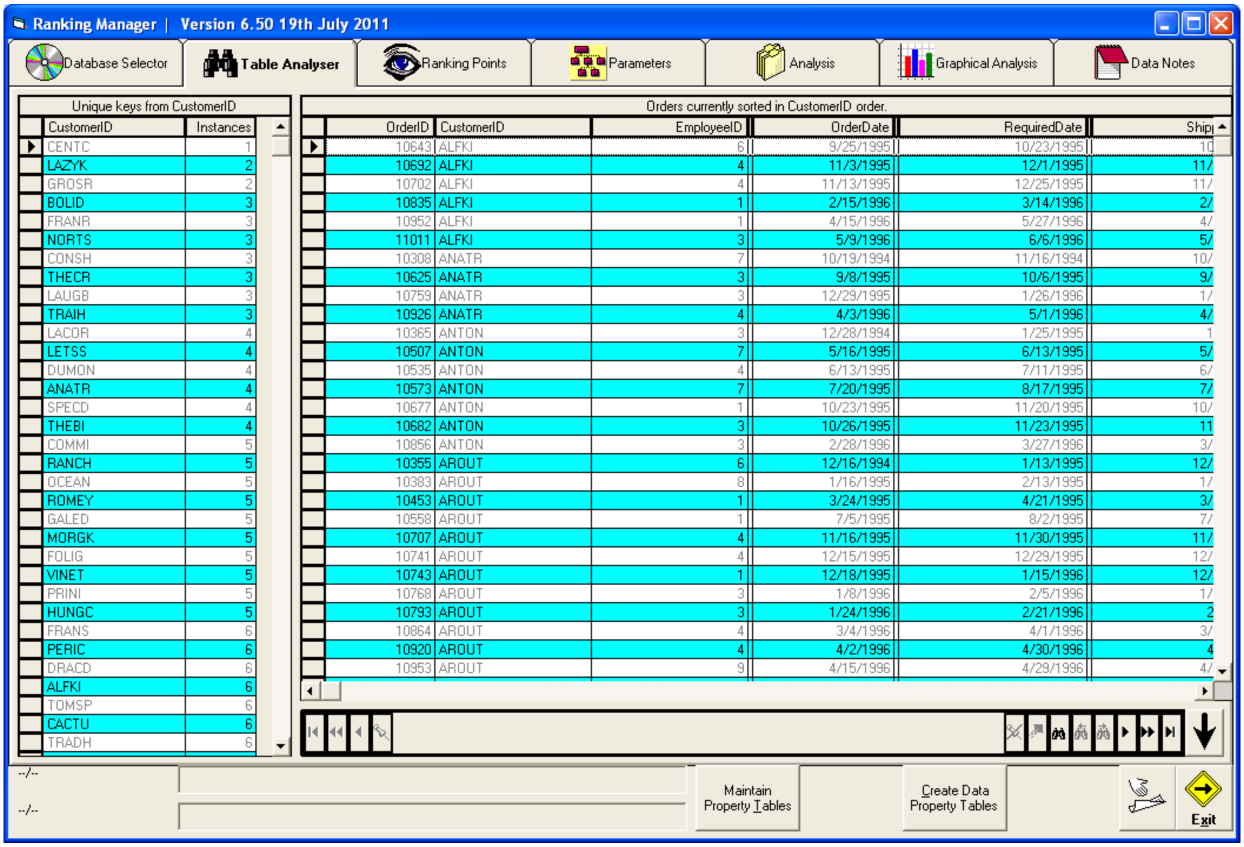
Iteration 5 – Experimental Trials
A key component of the experiment was the ability to introduce flags that indicated a level of data quality errors across the database. This iteration allowed the final modification to be introduced so that the data elements were 'cleaned' using different methods to compare and test the validity of ranked data as against unranked data.

5.1 Iteration 1 – Proof of Concept

5.1.1 *Observe*

This iteration was designed to show the concept that data elements could be ranked using a software artifact "Ranking Tool".

An initial component of the ranking tool was to display the target database as data manifestations such as tables, views, reports and forms.



Unique keys from CustomerID

CustomerID	Instances
CENTC	1
LAZYK	2
GROSR	2
BOLID	3
FRANR	3
NORTS	3
CONSH	3
THECR	3
LAUGB	3
TRAIH	3
LACOR	4
LETSS	4
DUMON	4
ANATR	4
SPECD	4
THEBI	4
COMMI	5
RANCH	5
OCEAN	5
ROMEY	5
GALED	5
MORGK	5
FOLIG	5
VINET	5
PRINI	5
HUNGC	5
FRANS	6
PERIC	6
DRACD	6
ALFKI	6
TOMSP	6
CACTU	6
TRADH	6

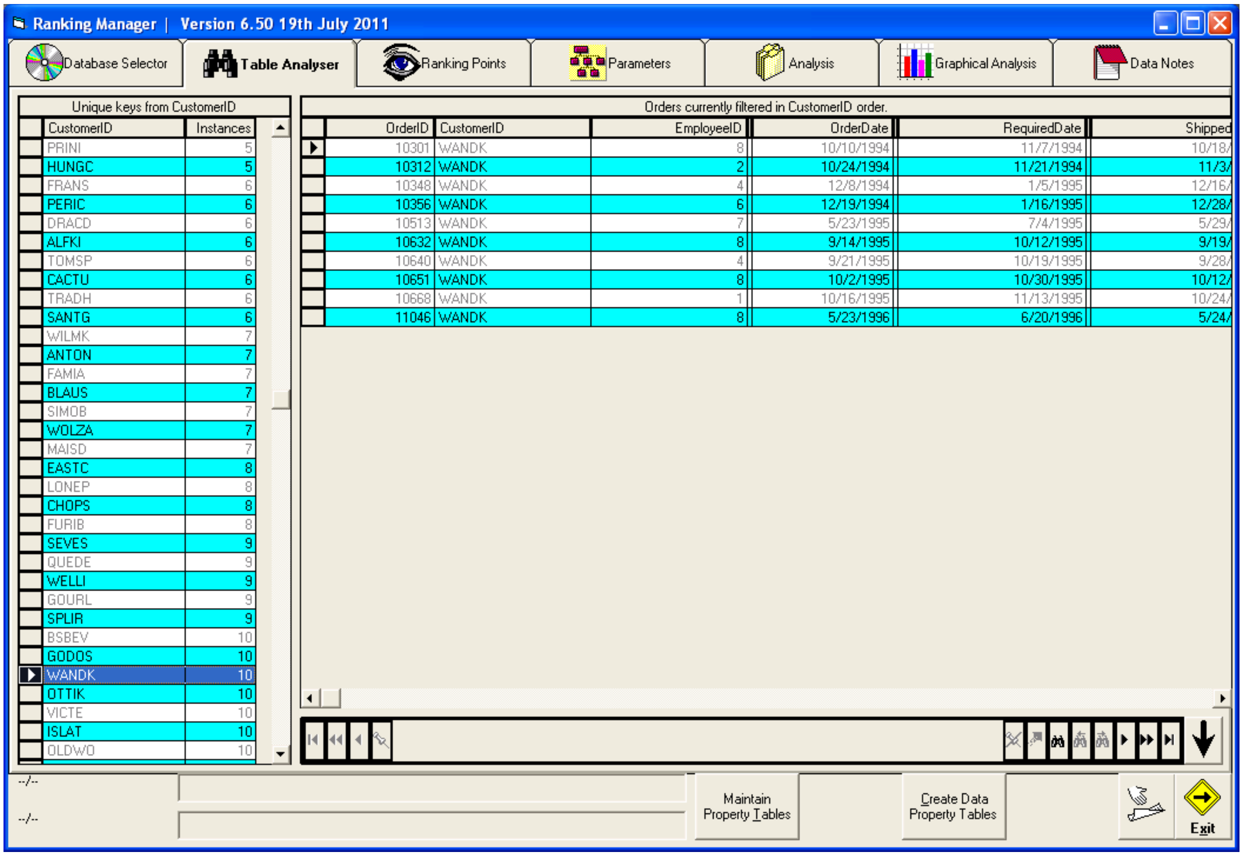
Orders currently sorted in CustomerID order.

OrderID	CustomerID	EmployeeID	OrderDate	RequiredDate	Ship
10643	ALFKI	6	9/25/1995	10/23/1995	10
10692	ALFKI	4	11/3/1995	12/1/1995	11/
10702	ALFKI	4	11/13/1995	12/25/1995	11/
10835	ALFKI	1	2/15/1996	3/14/1996	2/
10952	ALFKI	1	4/15/1996	5/27/1996	4/
11011	ALFKI	3	5/9/1996	6/6/1996	5/
10308	ANATR	7	10/19/1994	11/16/1994	10/
10625	ANATR	3	9/8/1995	10/6/1995	9/
10759	ANATR	3	12/29/1995	1/26/1996	1/
10926	ANATR	4	4/3/1996	5/1/1996	4/
10365	ANTON	3	12/28/1994	1/25/1995	1
10507	ANTON	7	5/16/1995	6/13/1995	5/
10535	ANTON	4	6/13/1995	7/11/1995	6/
10573	ANTON	7	7/20/1995	8/17/1995	7/
10677	ANTON	1	10/23/1995	11/20/1995	10/
10682	ANTON	3	10/26/1995	11/23/1995	11
10856	ANTON	3	2/28/1996	3/27/1996	3/
10355	AROUT	6	12/16/1994	1/13/1995	12/
10383	AROUT	8	1/16/1995	2/13/1995	1/
10453	AROUT	1	3/24/1995	4/21/1995	3/
10558	AROUT	1	7/5/1995	8/2/1995	7/
10707	AROUT	4	11/16/1995	11/30/1995	11/
10741	AROUT	4	12/15/1995	12/29/1995	12/
10743	AROUT	1	12/18/1995	1/15/1996	12/
10768	AROUT	3	1/8/1996	2/5/1996	1/
10793	AROUT	3	1/24/1996	2/21/1996	2
10864	AROUT	4	3/4/1996	4/1/1996	3/
10920	AROUT	4	4/2/1996	4/30/1996	4
10953	AROUT	9	4/15/1996	4/29/1996	4/

Figure 5-1 Database Calibration Management

Figure 5-1 Database Calibration Management displays the target database for any of the data manifestations in columns to facilitate sorting, searching, clustering by any key to facilitate testing for unique and related data for any data element. This component allows the target database calibration settings to be established.

The example shown in *Figure 5-2 Database Integrity Calibration* shows the Table:Orders with the sort key set to CustomerID. The left panel shows all unique instances of CustomerID and the related count of Table:Orders for each CustomerID.



Unique keys from CustomerID

CustomerID	Instances
PRINI	5
HUNGC	5
FRANS	6
PERIC	6
DRACD	6
ALFKI	6
TOMSP	6
CACTU	6
TRADH	6
SANTG	6
WILMK	7
ANTON	7
FAMIA	7
BLAUS	7
SIMOB	7
WOLZA	7
MAISD	7
EASTC	8
LONEP	8
CHOPS	8
FURIB	8
SEVES	9
QUEDE	9
WELLI	9
GOURL	9
SPLIR	9
BSBEV	10
GODOS	10
WANDK	10
OTTIK	10
VICTE	10
ISLAT	10
OLDWO	10

Orders currently filtered in CustomerID order.

OrderID	CustomerID	EmployeeID	OrderDate	RequiredDate	Shipped
10301	WANDK	8	10/10/1994	11/7/1994	10/18/
10312	WANDK	2	10/24/1994	11/21/1994	11/3/
10348	WANDK	4	12/8/1994	1/5/1995	12/16/
10356	WANDK	6	12/19/1994	1/16/1995	12/28/
10513	WANDK	7	5/23/1995	7/4/1995	5/29/
10632	WANDK	8	9/14/1995	10/12/1995	9/19/
10640	WANDK	4	9/21/1995	10/19/1995	9/28/
10651	WANDK	8	10/2/1995	10/30/1995	10/12/
10668	WANDK	1	10/16/1995	11/13/1995	10/24/
11046	WANDK	8	5/23/1996	6/20/1996	5/24/

Figure 5-2 Database Integrity Calibration

Figure 5-2 Database Integrity Calibration allows verification of the database relationships and associated integrity by allowing any displayed column to become a (temporary) primary index and a count of all related row data to be displayed. This verification establishes data element counts and relationships in all data manifestations (such as queries, report and data entry forms).

The analysis facility shown in *Figure 5-2 Database Integrity Calibration* allows individual data element tracking as they are created in tables through various ranking points.

Data element tracking demonstrates a reflection of the literature research that states that data can be used at the same time by many different users and can present different values for different users and uses.

As the data element components were examined during this process and some exclusions implemented. Excluded data included indexes, PIC (photographic) files and free text memo fields. The remaining data elements have been called "*contestable data elements*".

5.1.2 ***ANALYSE***

Examining the database, component tables, queries, reports and data entry screens showed that the tracking of data elements needed to be at row/column intersect as different data elements from the same column can participate in different and many views, reports and entry forms. They could also exhibit varying data quality errors. From the literature research, different data elements from the same column can be accessed by different users for different purposes and so acquire different value weightings.

For this reason, each row/column intersection (contestable data element) needed to be capable of accumulating rank-point values as it participated in various measurement points. Summation of the ranking point values for each data element represents the scalar value for each data element.

Provision was made in the artifact design to accommodate several databases with any number of tables. This research, however, was restricted to a single database.

5.1.3 ***TEST***

Various techniques were trialed to allow tracking against a number of ranking points for each data element. The North Wind Traders database was designed by Microsoft to have relational integrity set to 'ON'. A routine was designed and developed that selected each table in the database and then created a new ranking table where each row in the ranking table reflected each element in each row/column intersection from the North Wind Traders database.

Data elements that participated in relational integrity could not be logically ranked because corruption here would have rendered the relational database inoperable.

The expression '*contestable data elements*' was created to reflect data elements that were considered viable and contestable.

The ranking tool was changed to disallow non-contestable data elements and the routine re-run to check for the outcome.

Quality controls were developed as a series of cross tabulation displays that show the composition of all tables, views, data entry forms, and reports listed by their contributing table columns.

Each time the software is run it displays the following as verification (shown in *Figure 5-3 Ranking Point Values*) that the system is functioning as expected:

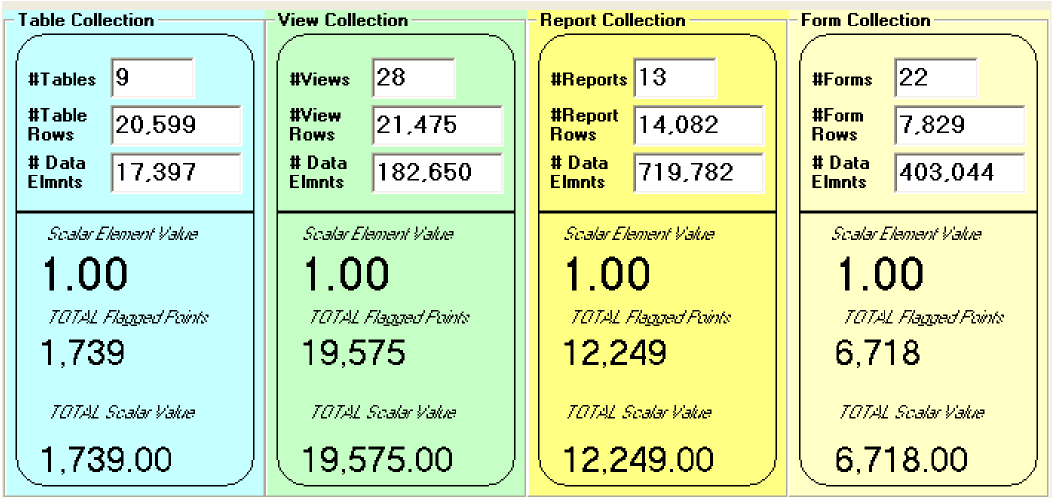


Table Collection

#Tables 9

#Table Rows 20,599

Data Elmnts 17,397

Scalar Element Value 1.00

TOTAL Flagged Points 1,739

TOTAL Scalar Value 1,739.00

View Collection

#Views 28

#View Rows 21,475

Data Elmnts 182,650

Scalar Element Value 1.00

TOTAL Flagged Points 19,575

TOTAL Scalar Value 19,575.00

Report Collection

#Reports 13

#Report Rows 14,082

Data Elmnts 719,782

Scalar Element Value 1.00

TOTAL Flagged Points 12,249

TOTAL Scalar Value 12,249.00

Form Collection

#Forms 22

#Form Rows 7,829

Data Elmnts 403,044

Scalar Element Value 1.00

TOTAL Flagged Points 6,718

TOTAL Scalar Value 6,718.00

Figure 5-3 Ranking Point Values

During this process, the ranking tool populates a Ranking_Table that stores the properties for each contestable data element.

Table 5-1 Initial Ranking Table Structure

Column Name	**Description**
Rank_Database	The name of the host database
Rank_*XXXX*Name	The name of the view, report or data entry form
Rank_ColumnName element	The column that presents the data
Rank_RowNumber	The row number of the table, view, report or form

In addition to the Ranking_Table, (see *Table 5-3*), a discrete table is generated to store detail for each contestable data element as it is subjected to each ranking point. This allows testing at each ranking point level as well as accumulation verification.

Initially, four ranking points were created as shown in *Table 5-2 Initial Ranking Points* :

Table 5-2 Initial Ranking Points

Tables
Views
Reports
Forms

Each ranking point data element is reflected in a discrete table as shown to allow the creation of a cross tabulation display to show correct data element tracking throughout the database:

Table 5-3 Aggregated Ranking Table

Name	**Type**	**Size**
Rank_Database	Text	30
Rank_TableName	Text	20
Rank_FieldType	Text	20
Rank_Table	**Single**	**4**
Rank_View	**Single**	**4**
Rank_Report	**Single**	**4**
Rank_Form	**Single**	**4**

The bold font indicates the additional information that extension each of the data element properties. The additional data manifestations are represented as fields Rank_Table, Rank_View, Rank_Report and Rank_Form that store the aggregated ranking value for each contestable data element as it is recognised in each data manifestation.

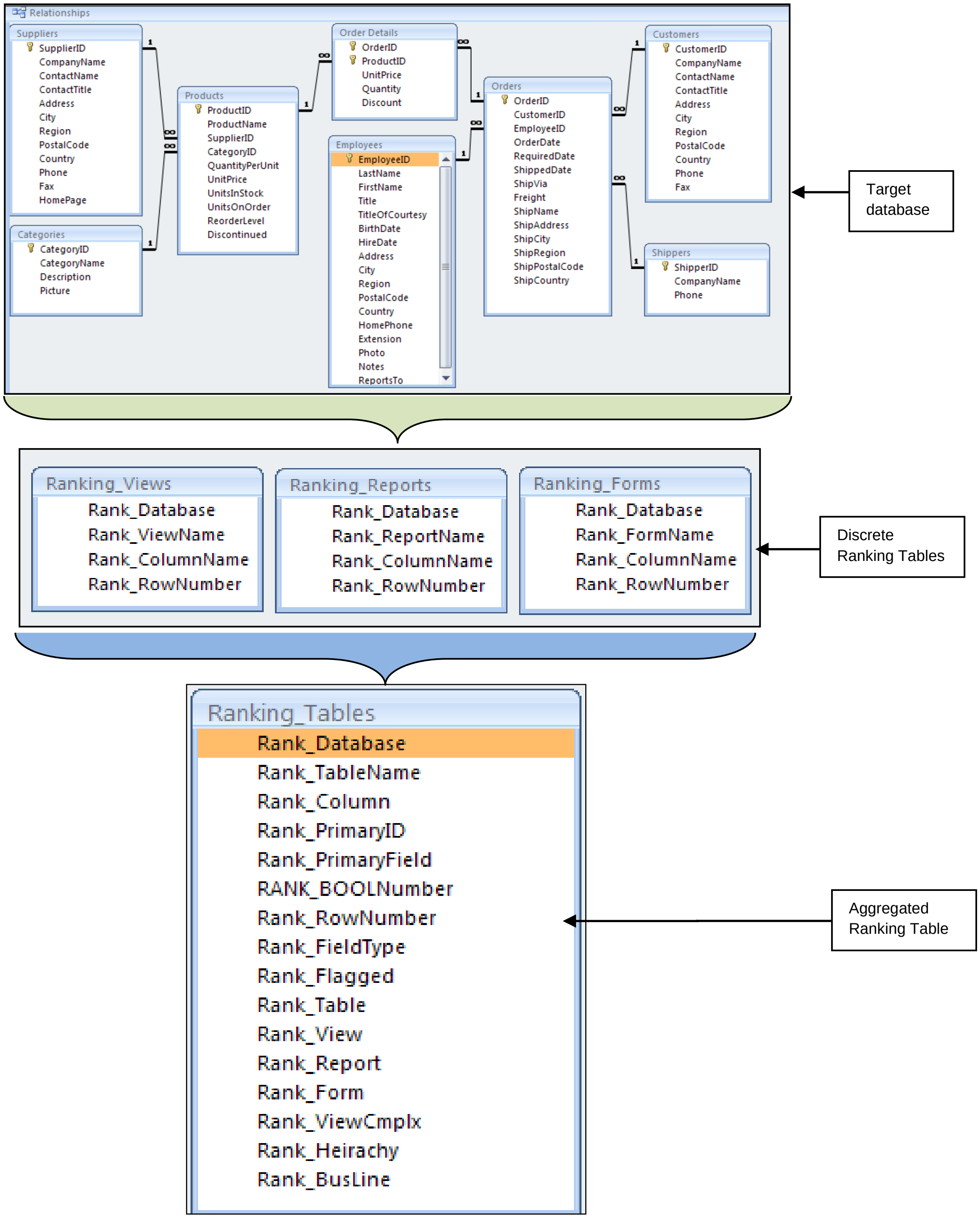


Figure 5-4 Ranking Point Properties / Database Relationship

The relationship between the target database, discrete ranking tables and the aggregated ranking table is shown in Figure 5-4 Ranking Point Properties / Database Relationship.

5.1.4 *CROSS TABULATION TABLES (X/TAB)*

Cross Tabulations (X/TAB) were created from the discrete ranking tables and the aggregated ranking table as a configuration mechanism to verify successful contestable data elements tracking from each of the four data manifestations: Tables, Views, Data Entry Forms and Reports.

5.1.5 *X/TAB - TABLES*

*Figure 5-5 Table-Based Contestable Data Elements as an X/*TAB shows the processed contestable data elements from all seven tables in the database. Each column shows the table names and the rows show the table column names.

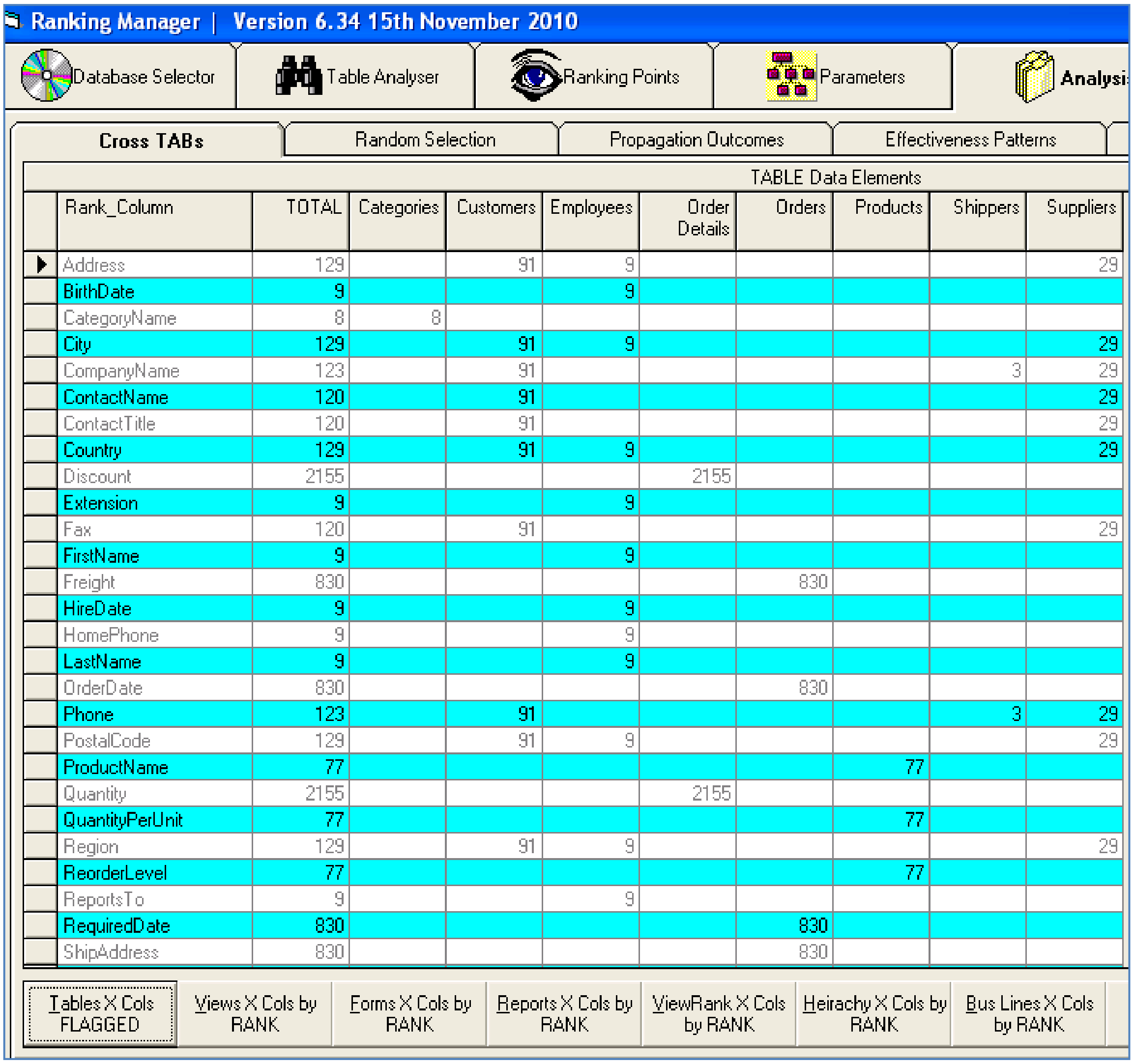


			TABLE Data Elements						
Rank_Column	TOTAL	Categories	Customers	Employees	Order Details	Orders	Products	Shippers	Suppliers
Address	129		91	9					29
BirthDate	9			9					
CategoryName	8	8							
City	129		91	9					29
CompanyName	123		91					3	29
ContactName	120		91						29
ContactTitle	120		91						29
Country	129		91	9					29
Discount	2155				2155				
Extension	9			9					
Fax	120		91						29
FirstName	9			9					
Freight	830					830			
HireDate	9			9					
HomePhone	9			9					
LastName	9			9					
OrderDate	830					830			
Phone	123		91					3	29
PostalCode	129		91	9					29
ProductName	77						77		
Quantity	2155				2155				
QuantityPerUnit	77						77		
Region	129		91	9					29
ReorderLevel	77						77		
ReportsTo	9			9					
RequiredDate	830					830			
ShipAddress	830					830			

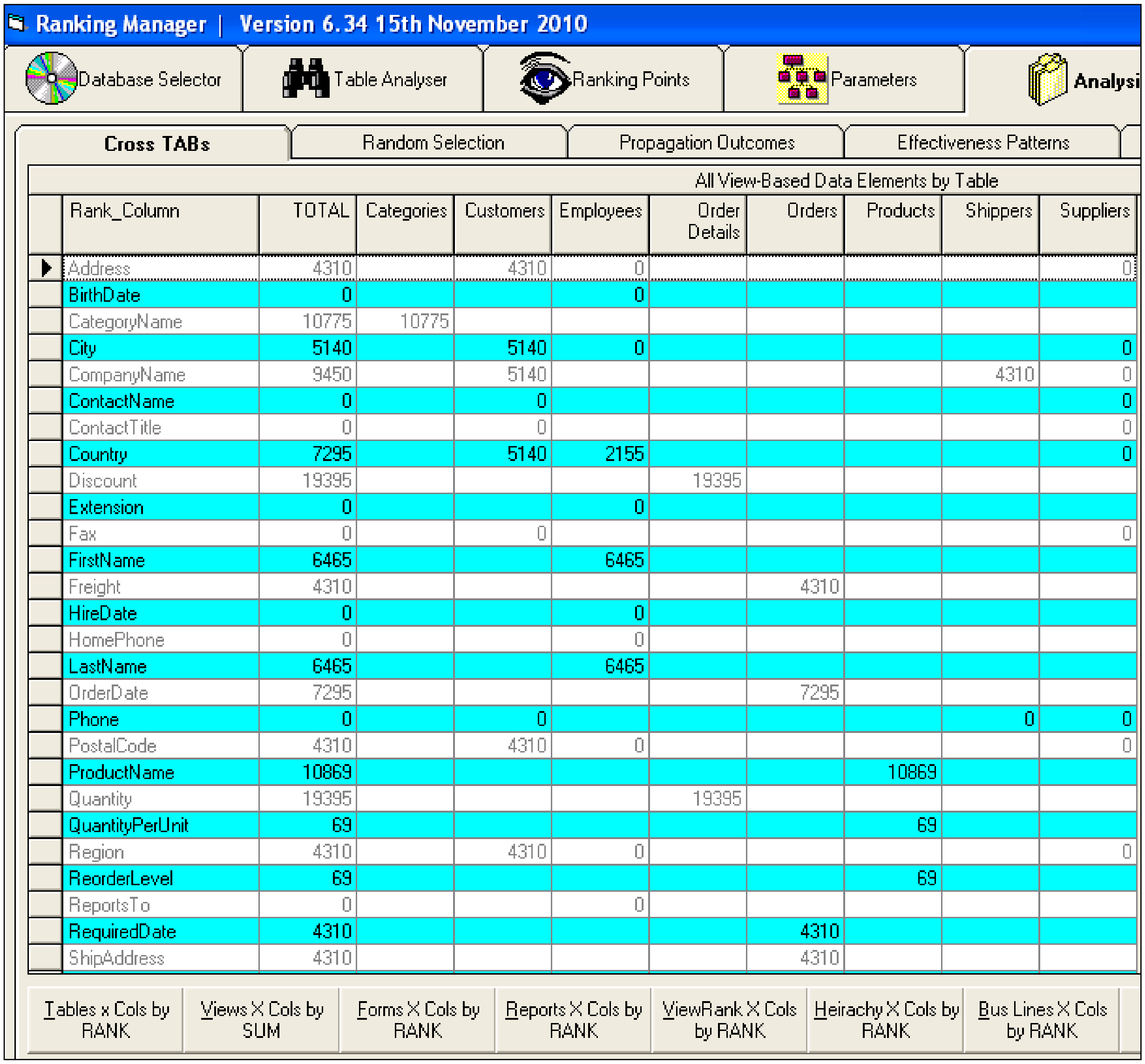
Figure 5-5 Table-Based Contestable Data Elements as an X/TAB

The counts for each contestable data element are displayed at the row/column intersection. This demonstrates that all expected contestable data elements have been accounted for and processed.

Each row /column intersection is treated as a discrete data element. Each contestable data element in each table are assigned a ranking value of 1 and unique key that allows it to be tracked as it is encountered in each data manifestation.

5.1.6 *X/TAB - VIEWS*

Figure 5-6 View-Based Contestable Data Element as an X/TAB shows the processed contestable data elements from all views by tables in the database. Each column shows the table names and the rows show the table column names.



All View-Based Data Elements by Table

Rank_Column	TOTAL	Categories	Customers	Employees	Order Details	Orders	Products	Shippers	Suppliers
Address	4310		4310	0					0
BirthDate	0			0					
CategoryName	10775	10775							
City	5140		5140	0					0
CompanyName	9450		5140					4310	0
ContactName	0		0						0
ContactTitle	0		0						0
Country	7295		5140	2155					0
Discount	19395				19395				
Extension	0			0					
Fax	0		0						0
FirstName	6465			6465					
Freight	4310					4310			
HireDate	0			0					
HomePhone	0			0					
LastName	6465			6465					
OrderDate	7295					7295			
Phone	0		0					0	0
PostalCode	4310		4310	0					0
ProductName	10869						10869		
Quantity	19395				19395				
QuantityPerUnit	69						69		
Region	4310		4310	0					0
ReorderLevel	69						69		
ReportsTo	0			0					
RequiredDate	4310					4310			
ShipAddress	4310					4310			

Figure 5-6 View-Based Contestable Data Element as an X/TAB

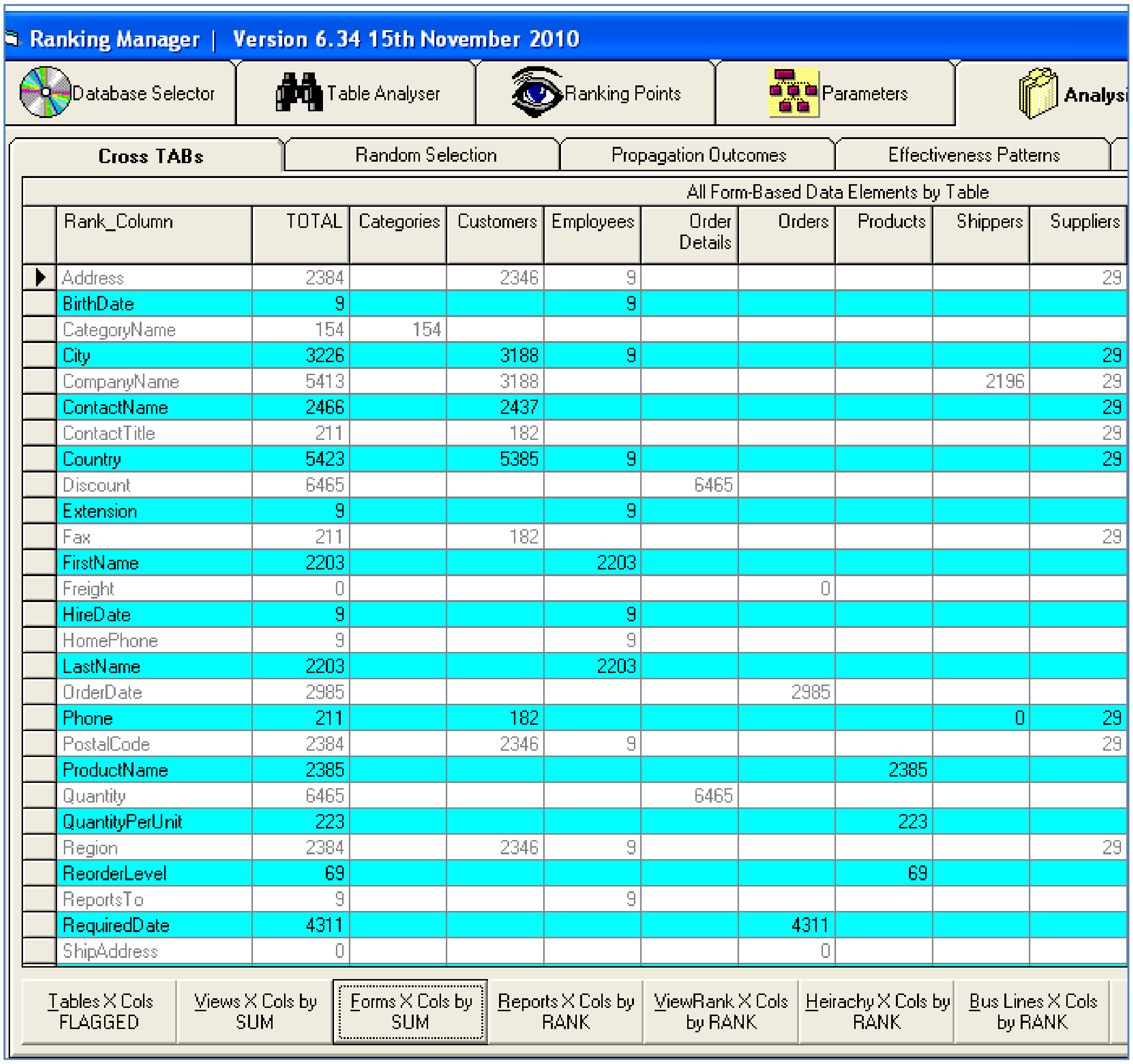
The counts for each contestable data element are displayed at the row/column intersection. This demonstrates that all expected contestable data elements have been accounted for and processed.

Views can present many contestable data elements that participate in one or more views as some data elements will participate in several different data aggregations to suit varying business requirements.

Contestable data elements are scored each time they are detected in and data manifestation. These scored are an aggregation of the scores from tables, views, reports and data forms.
\
Frequently accessed data elements are therefore ranked higher than less frequently used data elements.

5.1.7 *X/TAB - DATA ENTRY FORMS*

Figure 5-7 Data Entry Form-Based Contestable Data Elements as an X/TAB shows the processed contestable data elements from all Data Forms by tables in the database. Each column shows the table names and the rows show the table column names. The counts for each contestable data element are displayed at the row/column intersection. This demonstrates that all expected contestable data elements have been accounted for and processed.



	All Form-Based Data Elements by Table								
Rank_Column	TOTAL	Categories	Customers	Employees	Order Details	Orders	Products	Shippers	Suppliers
Address	2384		2346	9					29
BirthDate	9			9					
CategoryName	154	154							
City	3226		3188	9					29
CompanyName	5413		3188					2196	29
ContactName	2466		2437						29
ContactTitle	211		182						29
Country	5423		5385	9					29
Discount	6465				6465				
Extension	9			9					
Fax	211		182						29
FirstName	2203			2203					
Freight	0					0			
HireDate	9			9					
HomePhone	9			9					
LastName	2203			2203					
OrderDate	2985					2985			
Phone	211		182					0	29
PostalCode	2384		2346	9					29
ProductName	2385						2385		
Quantity	6465				6465				
QuantityPerUnit	223						223		
Region	2384		2346	9					29
ReorderLevel	69						69		
ReportsTo	9			9					
RequiredDate	4311					4311			
ShipAddress	0					0			

Figure 5-7 Data Entry Form-Based Contestable Data Elements as an X/TAB

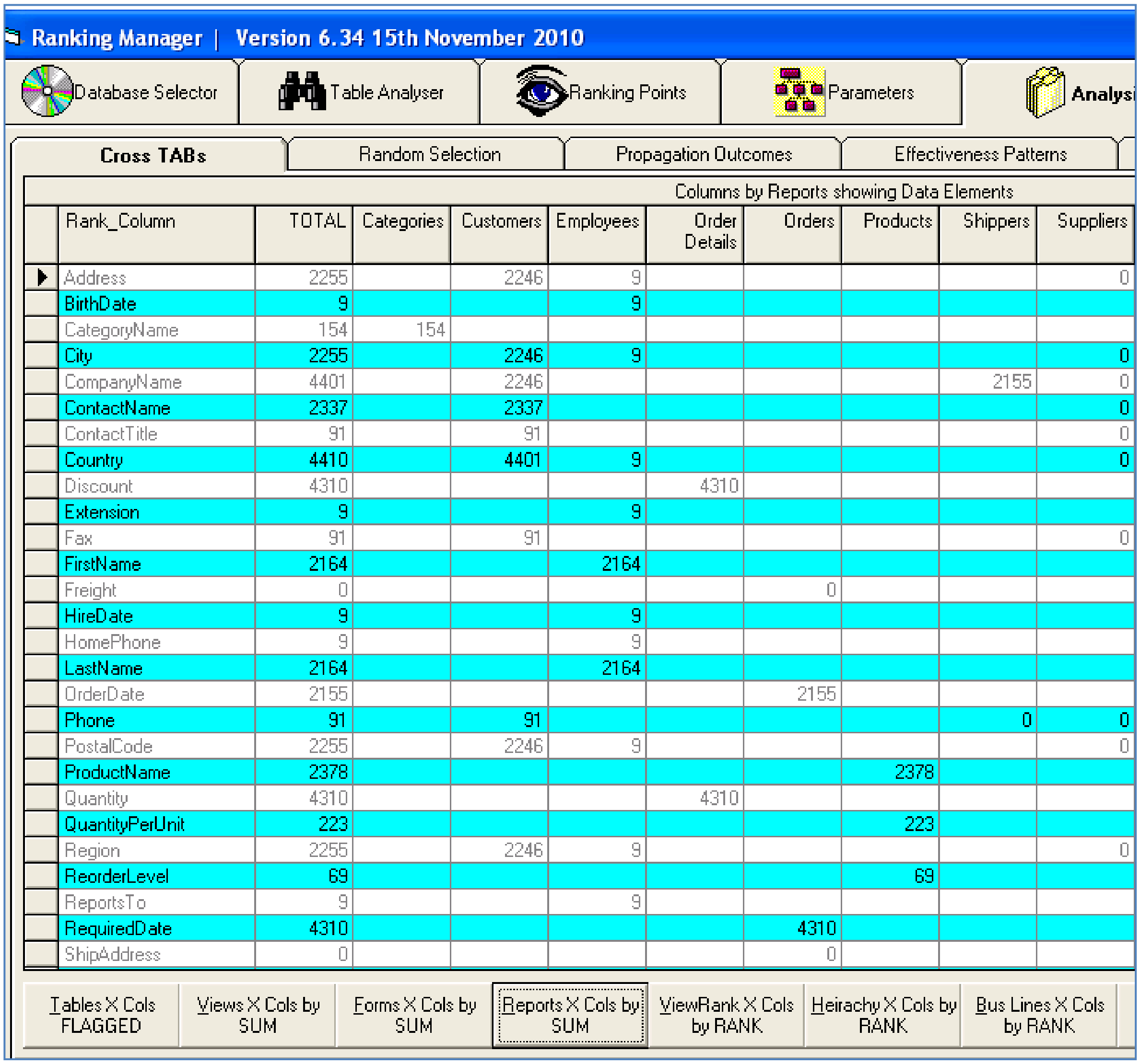
Contestable data elements participating in forms are scored each time they are detected in a view.   These scored are an aggregation of the scores from tables, views, report and data forms.

5.1.8 *X/TAB - REPORTS*

Figure 5-7 Data Entry Form-Based Contestable Data Elements as shows the contestable data elements from all Reports by tables in the database.

Each column shows the table names and the rows show the table column names. The counts for each contestable data element are displayed at the row/column intersection. This demonstrates that all expected contestable data elements have been accounted for and processed.

Note that as for views and forms, some contestable data elements participate in many reports.



Columns by Reports showing Data Elements

Rank_Column	TOTAL	Categories	Customers	Employees	Order Details	Orders	Products	Shippers	Suppliers
Address	2255		2246	9					0
BirthDate	9			9					
CategoryName	154	154							
City	2255		2246	9					0
CompanyName	4401		2246					2155	0
ContactName	2337		2337						0
ContactTitle	91		91						0
Country	4410		4401	9					0
Discount	4310				4310				
Extension	9			9					
Fax	91		91						0
FirstName	2164			2164					
Freight	0					0			
HireDate	9			9					
HomePhone	9			9					
LastName	2164			2164					
OrderDate	2155					2155			
Phone	91		91					0	0
PostalCode	2255		2246	9					0
ProductName	2378						2378		
Quantity	4310				4310				
QuantityPerUnit	223						223		
Region	2255		2246	9					0
ReorderLevel	69						69		
ReportsTo	9			9					
RequiredDate	4310					4310			
ShipAddress	0					0			

Figure 5-8 Report Contestable Data Elements as an X/TAB

5.1.9 *ITERATION 1 - CONCLUSION*

This iteration achieved the initial proof-of-concept and so prepared the software tool for further analysis and development to address the question around assigning values to contestable data elements in a database.

The approach using cross tabulation tables to display the outcome and relationship of table data elements against tables, views, data entry forms and reports allowed processing verification for the software. The propagation of contestable database elements as they are used in views and reports presents a database profile. This propagation illustrates how data quality errors can similarly propagate throughout a database.

5.2 ITERATION 2 – INITIAL MODIFICATION

5.2.1 *OBSERVATION*

Having observed and verified the output from iteration 1, modifications were planned, introduced and executed for the second iteration.

The initial approach using the single vector scaling method from (Page et al., 1998) did not allow variable weightings to be introduced as various ranking points. Given that the literature suggests that different users may make different decisions based on the same information. It is also assumed that higher ranked users in an organisation are more likely to make more important decisions. 2.7, Ranking Approaches, p24

Each of the four ranking points added a standard value to each contestable data element each time they were detected in each ranking point. This approach was sufficient to demonstrate the process; but did not reflect the literature research that described adequate data quality is also a function of the users as well as usage. The model reflected technical weighting, but did not reflect business usage or 'importance' weighting. See 2.7, Ranking Approaches, p24 for literature describing weighting research and 2.3, The Changing Perceptions of Data Quality, p12 describing how data quality perceptions have evolved top reflect business users and usage.

Ranking approaches described in 2.7, Ranking Approaches, Page 24 were modified and combined to develop a reflection of the ranking points used by the ranking engine to calculate and accumulate the scalar data values for each contestable data element.

The Page Rank© ranking system (See 2.7, Ranking Approaches, p24) was modified with an "*intrinsic manifold structure*" that reflects a weighted ranking based on a '*network relationship*' that "*specifies the relative contributions to the ranking scores from neighbors*".

This modification has been further modified by assigning scalar value aggregation that ranks contestable data elements based on the "*network nodes*" traversed.

This modification is reflected with the introduction of variable structure ranking biases at each of the ranking points.

Each of the ranking points has configurable options to allow variability within the ranking engine to allow different numeric weights to be created (or removed) at each ranking point.

These variable ranking points offer the opportunity for further research when applied to operational databases in an organisation.

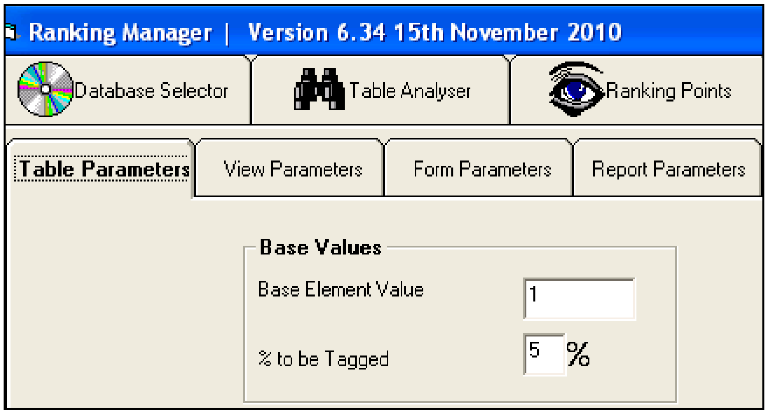
5.2.2 *Analyse*

The ranking tool at this stage was operational and was further enhanced to allow variability within each ranking point. The variability reflects the business usage as well as technical ranking as described in 2.7, Ranking Approaches, p24

Ranking Point 1

The Base Element Value can be changed but predictably offered no material effect on the relative value of the scaled data elements. This ranking point could be modified in the future to offer variable commencement values for different types of data elements.

The percentage to be tagged is a facility that allows data quality errors to be simulated at contestable data element level in the tables. See 5.5, Iteration 5 – Experimental Implementation and Trials, p71 for details about this facility.



Ranking Manager | Version 6.34 15th November 2010

Base Values

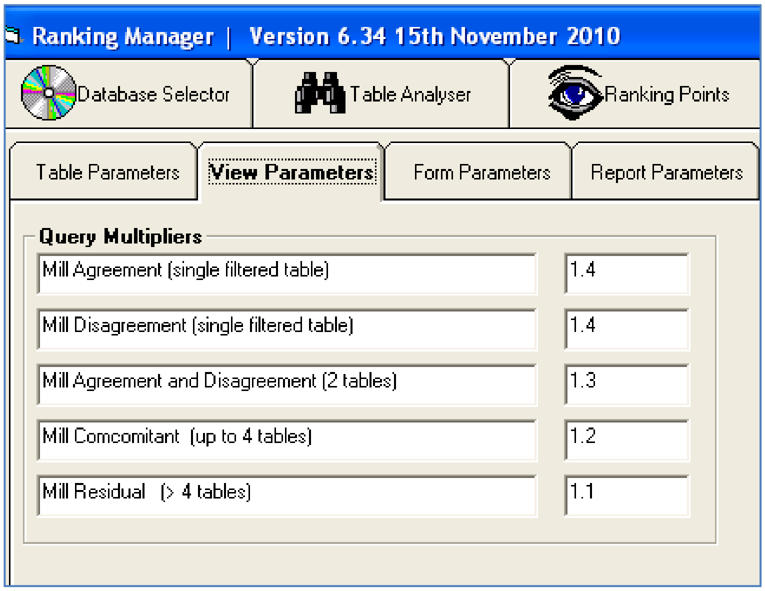
Base Element Value: 1

% to be Tagged: 5 %

Figure 5-9 Table Element Parameters

Ranking Point 2

This ranking point was enhanced to reflect the information cannons proposed by John Stuart Mill (See 2.7.1, *Technical Weighting*, p26). There was no literature found that describes view weighting based on information classification. This ranking addition classifies views into a set of data facets ("query multipliers") that reflect the value data elements that contributed to the views.



Query Multipliers

Mill Agreement (single filtered table) 1.4

Mill Disagreement (single filtered table) 1.4

Mill Agreement and Disagreement (2 tables) 1.3

Mill Comcomitant (up to 4 tables) 1.2

Mill Residual (> 4 tables) 1.1

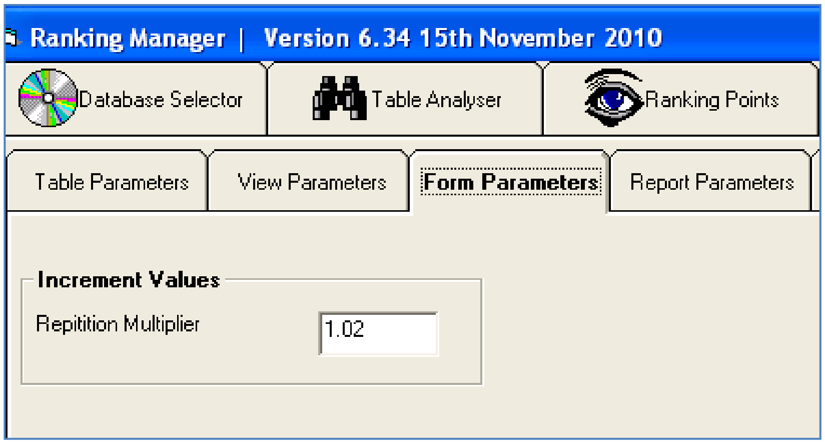
Figure 5-10 Ranking View Parameters (Data Facets)

The data facets illustrated in *Figure 5-10 Ranking View Parameters* show that they can be modified to offer different weightings for each facet. This option allows further research into this facility when analysing different databases.

Ranking Point 3

The data form ranking point was enhanced to reflect the notion that data elements that contributed to a data entry form would present some level of added significance in that much of the data used to populate a form is contributing to data collection.

The rationale is that form-based data contributes to new data rows in tables. Therefore flawed data would propagate data quality errors through data entry forms. No literature was found that describes this issue.



Increment Values

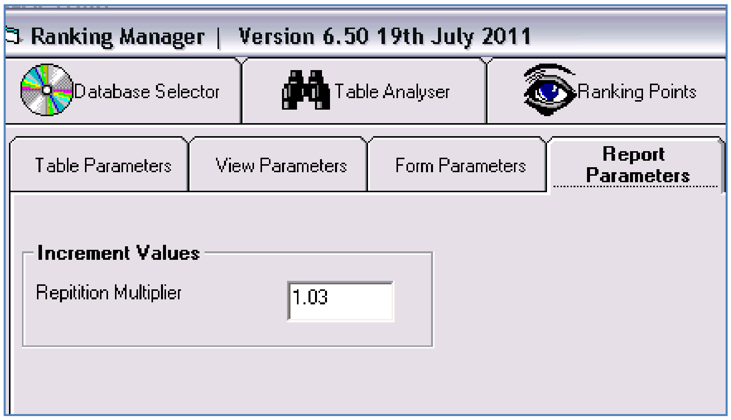
Repitition Multiplier 1.02

Figure 5-11 Form Element Parameters

This facility too could be enhanced for further research when conducting evaluations against live databases.

Ranking Point 4

This ranking point was enhanced to rank data elements that contributed to a report as they would present some level of significance in that much of the data used to populate a report would be contributing to decision making. Should any of these contributing data elements prove flawed, then the flawed data would contribute to poor decision making (<u>See 2.1, Data Quality and Decision making, p7</u>).



Increment Values

Repitition Multiplier 1.03

Figure 5-12 Report Element Parameters

Although the effect of poor data quality on effective decision making is well documented, there is no literature that reflects a ranking approach that weights data when it is used for reports.

Figure 5-13 Ranking Measurement Points – 2nd Iteration models the initial 4 ranking points in database context:

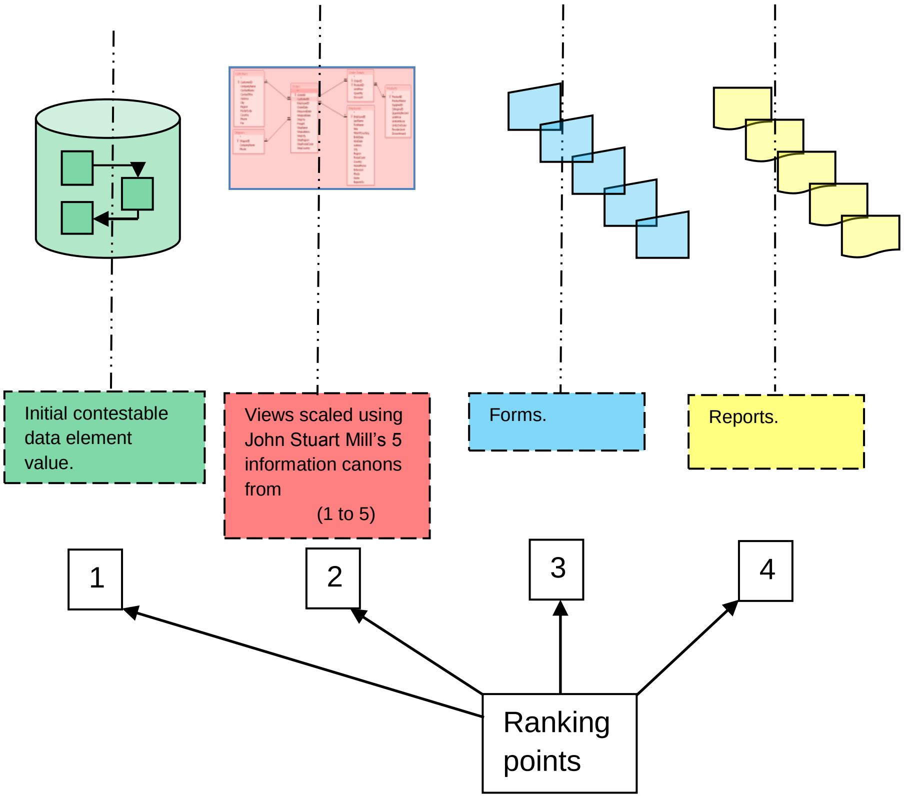

Initial contestable data element value.
Views scaled using John Stuart Mill's 5 information canons from (1 to 5)
Forms.
Reports.
1
2
3
4
Ranking points

Figure 5-13 Ranking Measurement Points – 2nd Iteration

Each of the ranking points has been described in: *Figure 5-9 Table Element Parameters*, *Figure 5-10 Ranking View Parameters (Data Facets)*, *Figure 5-11 Form Element Parameters*, and *Figure 5-12 Report Element Parameters.*

5.2.3 *EXPERIMENT*

The artifact was tested using various values at each of the ranking points and a correlation of co-efficient generated to show the relationship strength between each of the ranking points.

Although the weighted values introduced did not affect the correlation strengths significantly, the values introduced in the 5.2.4, *Test*, p63 proved to be optimal for correlation values. Given that the contestable data values are scalar rather than intrinsic, this outcome demonstrates the concept.

5.2.4 *TEST*

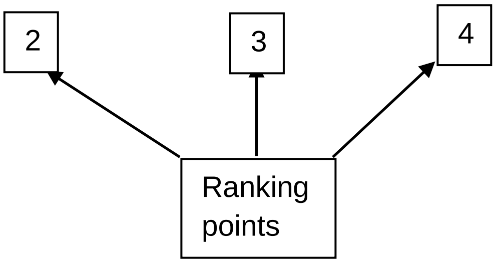
A low correlation (less than .3), would show that the weightings and the ranking points would have lost much relevance.

A table-based correlation matrix was created to test the correlation between the ranking points. The sample size is the complete set of contestable data elements (n=N). The correlation (Spearman's Correlation of co-efficiency) shows that the relationship between the ranking points to be significant given the population of 17397.

This table shows however that the relationships between the ranking points are high with correlations ranging from 0.999125 to 0.715360. This outcome confirms that there is a strong relationship between each of the ranking points.

Table 5-4 TEST: Correlation of Coefficient for N Elements for Three Ranking Points

	View	Report	Form
View	1.000000	0.720802	0.715360
Report	0.720802	1.000000	0.995633
Form	0.715360	0.995633	1.000000



Tests were conducted by neutralising the values in the each of the ranking points (setting the scaling factor to 1.00) and the difference offered a slightly lower correlation showing that the weightings offer some calibration advantage.

Note also that the table design shows the same correlation between (for example) "Form to Report" and "Report to Form". These two correlations are identical suggesting that the software is operating correctly.

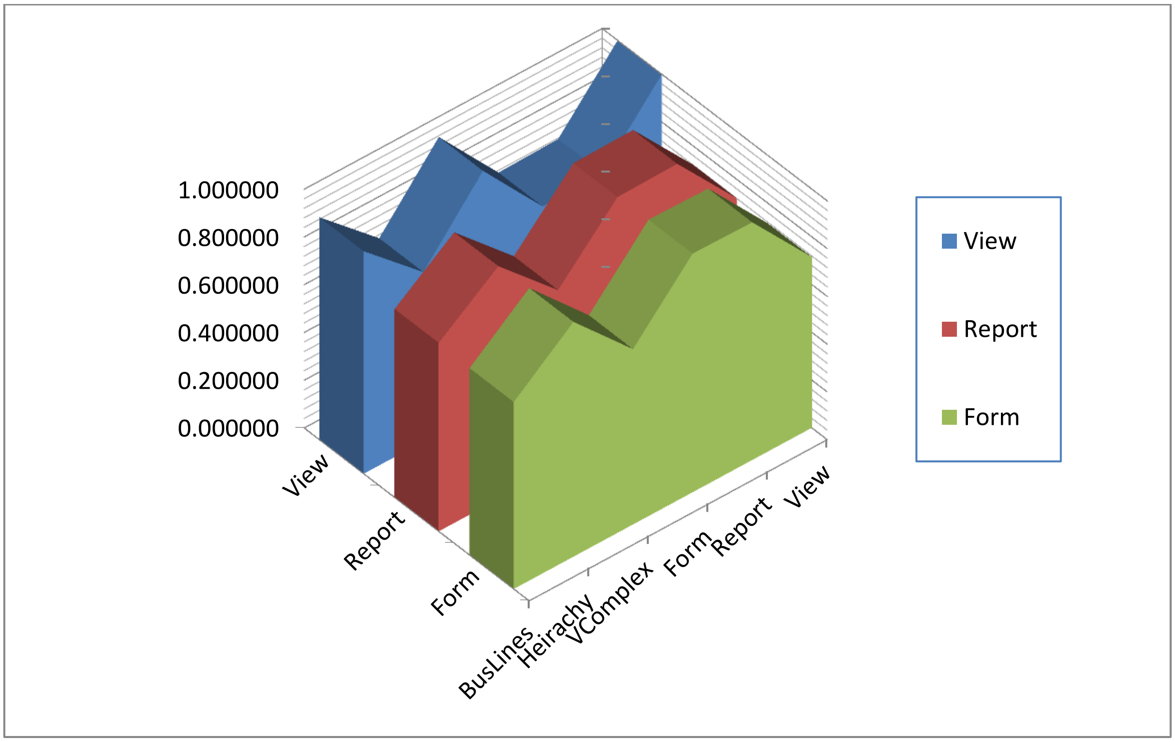


Figure 5-14 Correlation Relationship between Four Ranking Points

5.2.5 *ITERATION 2 - CONCLUSION*

The ranking points and the associated values are significant and demonstrate the notion of ranking the contestable data elements that offers a scalar value that is consistent across the ranking points in the database. This process also demonstrates a solution to the issues raised in the ranking research described in 2.7, *Ranking Approaches*, p24.

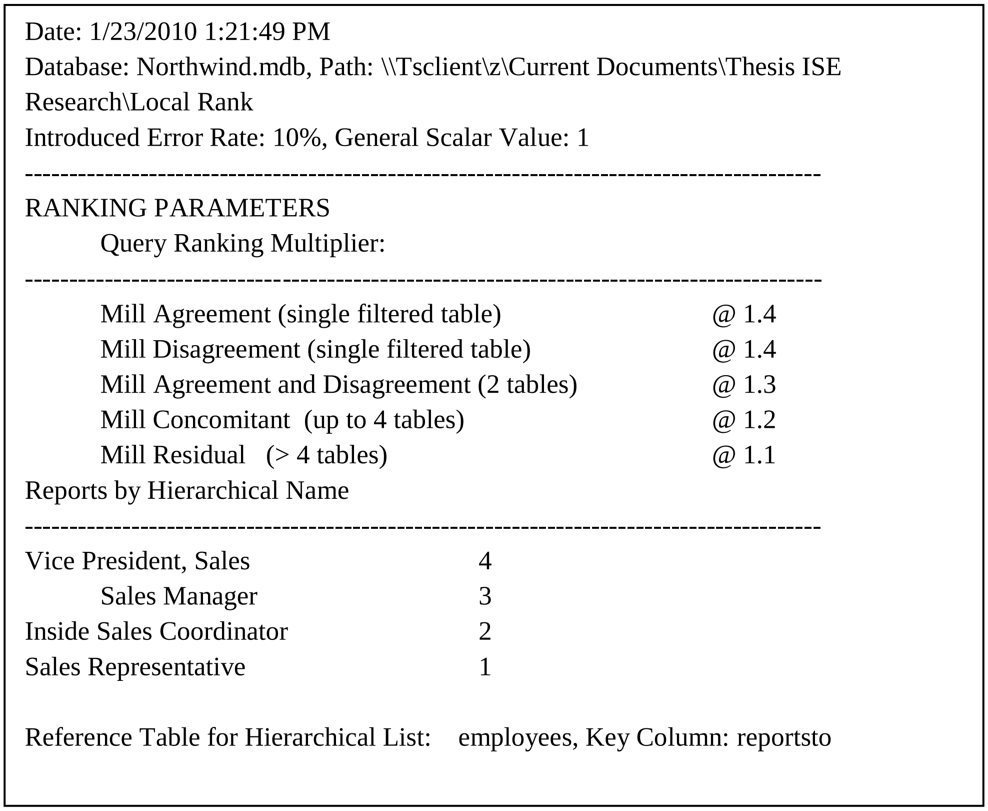
5.3 Iteration 3 - Ranking Point Calibration

5.3.1 *Observation*

A key requirement became apparent that was the need to log all executions of the ranking tool automatically so that records could be maintained showing the results of each run. The log table was expanded as additional facilities were added to the ranking tool.

As the iterations continued, base-line results were also included in the log table making the history of each run easily comparable.

A portion of the log table is illustrated in
Figure 5-15 Session-based Log File Sample*.*



```
Date: 1/23/2010 1:21:49 PM
Database: Northwind.mdb, Path: \\Tsclient\z\Current Documents\Thesis ISE
Research\Local Rank
Introduced Error Rate: 10%, General Scalar Value: 1
-------------------------------------------------------------------------------------------
RANKING PARAMETERS
        Query Ranking Multiplier:
-------------------------------------------------------------------------------------------
        Mill Agreement (single filtered table)                  @ 1.4
        Mill Disagreement (single filtered table)               @ 1.4
        Mill Agreement and Disagreement (2 tables)              @ 1.3
        Mill Concomitant  (up to 4 tables)                      @ 1.2
        Mill Residual   (> 4 tables)                            @ 1.1
Reports by Hierarchical Name
-------------------------------------------------------------------------------------------
Vice President, Sales                   4
        Sales Manager                   3
Inside Sales Coordinator                2
Sales Representative                    1

Reference Table for Hierarchical List:   employees, Key Column: reportsto
```

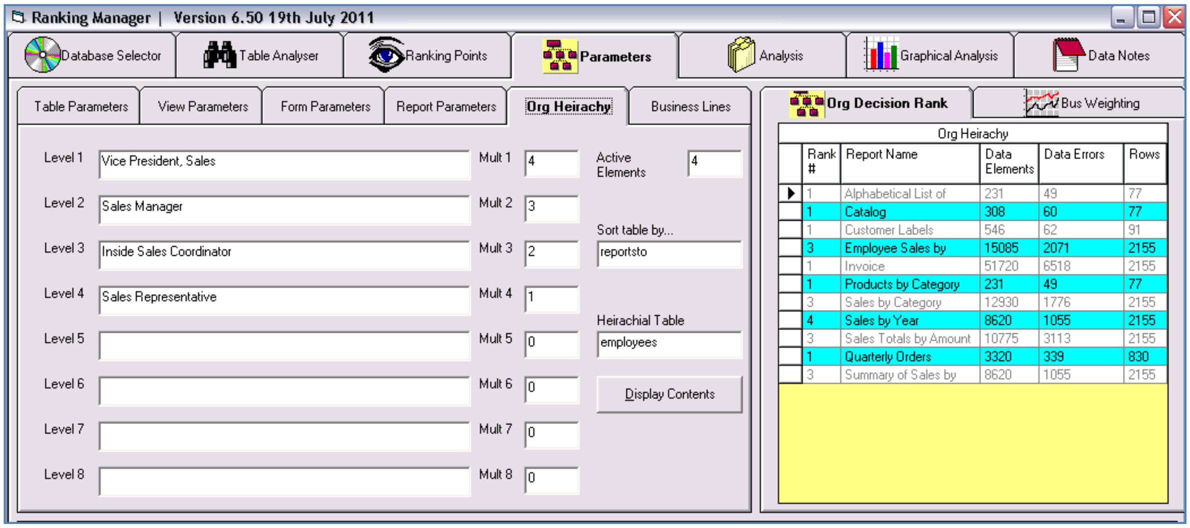
Figure 5-15 Session-based Log File Sample

This approach shown in
Figure 5-15 Session-based Log File Sample allows comparisons to be made for various runs against different variables in the system as well as a record of all experiments conducted against the target database.

Ranking Points

In addition to the ranking points described in iteration 2, an additional report-based ranking point was introduced that reflects the hierarchical position of the decision maker. The assumption here is that the more significant a decision-maker is in an organisation, the more significant their decision making will be.

The report ranking follows the “reports to” column in Table_Employees to allow hierarchical ranking of the decision-maker and therefore to add scalar value to the contestable data elements that contribute to these reports.



Level 1	Vice President, Sales	Mult 1	4
Level 2	Sales Manager	Mult 2	3
Level 3	Inside Sales Coordinator	Mult 3	2
Level 4	Sales Representative	Mult 4	1
Level 5		Mult 5	0
Level 6		Mult 6	0
Level 7		Mult 7	0
Level 8		Mult 8	0

Active Elements: 4

Sort table by...: reportsto

Heirachial Table: employees

Org Heirachy

Rank #	Report Name	Data Elements	Data Errors	Rows
1	Alphabetical List of	231	49	77
1	Catalog	308	60	77
1	Customer Labels	546	62	91
3	Employee Sales by	15085	2071	2155
1	Invoice	51720	6518	2155
1	Products by Category	231	49	77
3	Sales by Category	12930	1776	2155
4	Sales by Year	8620	1055	2155
3	Sales Totals by Amount	10775	3113	2155
1	Quarterly Orders	3320	339	830
3	Summary of Sales by	8620	1055	2155

Figure 5-16 'Reports To' Ranking Point

Figure 5-13 Ranking Measurement Points – 2nd Iteration shows the process for weighting hierarchical values based on organisational position. The reports have been listed showing the primary users of these reports.

The fourth column in the table shown in *Figure 5-13 Ranking Measurement Points – 2nd Iteration* shows the number of 'seeded' data quality errors (at table manifestation level) at 10% across the report selection.

This method now offers a five ranking point set:

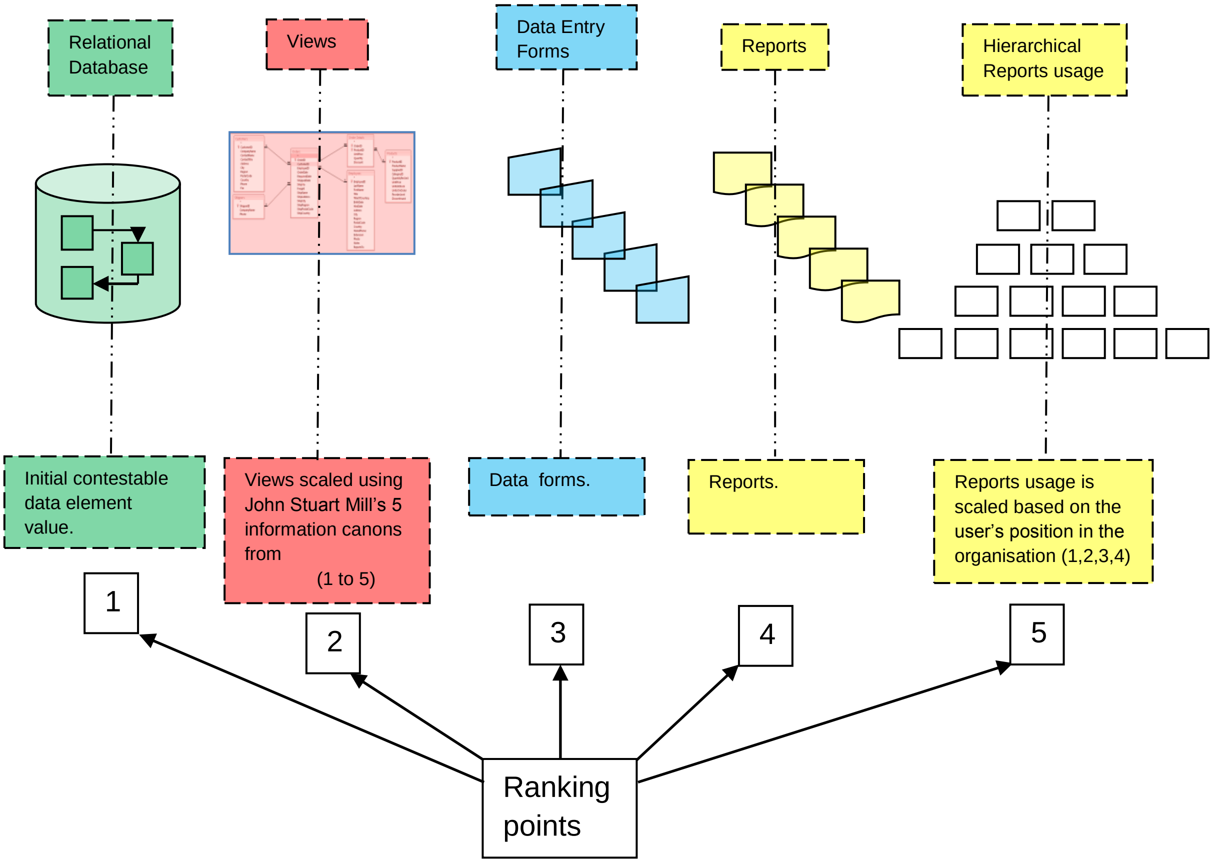

Relational Database
Views
Data Entry Forms
Reports
Hierarchical Reports usage
Initial contestable data element value.
Views scaled using John Stuart Mill's 5 information canons from (1 to 5)
Data forms.
Reports.
Reports usage is scaled based on the user's position in the organisation (1,2,3,4)
1
2
3
4
5
Ranking points

Figure 5-17 Five Ranking Measurement Points - 3rd Iteration

The correlation grid (*Table 5-5*) was regenerated for the five ranking points and the results observed.

The results for the correlation test show that the new ranking point (*Figure 5-13 Ranking Measurement Points – 2nd Iteration*) is highly significant when correlated with the first 4 ranking points.

Table 5-5 TEST: Correlation of Coefficient for N Elements for Five Ranking Points

	View	**Report**	**Form**	**View Ranking**	**Hierarchy**
View	1.000000	0.720802	0.715360	0.999125	0.709555
Report	0.720802	1.000000	0.995633	0.740439	0.975565
Form	0.715360	0.995633	1.000000	0.734334	0.983197
View Ranking	0.999125	0.740438	0.734334	1.000000	0.725979
Hierarchy	0.709555	0.975565	0.983197	0.725979	1.000000

The correlation table above shows that the relationship between Views, Reports, Forms, View Complexity and Hierarchical order is highly significant.   The yellow highlights are a control to verify that the Spearman's correlation formula is being applied correctly.

5.3.2 *ANALYSE*

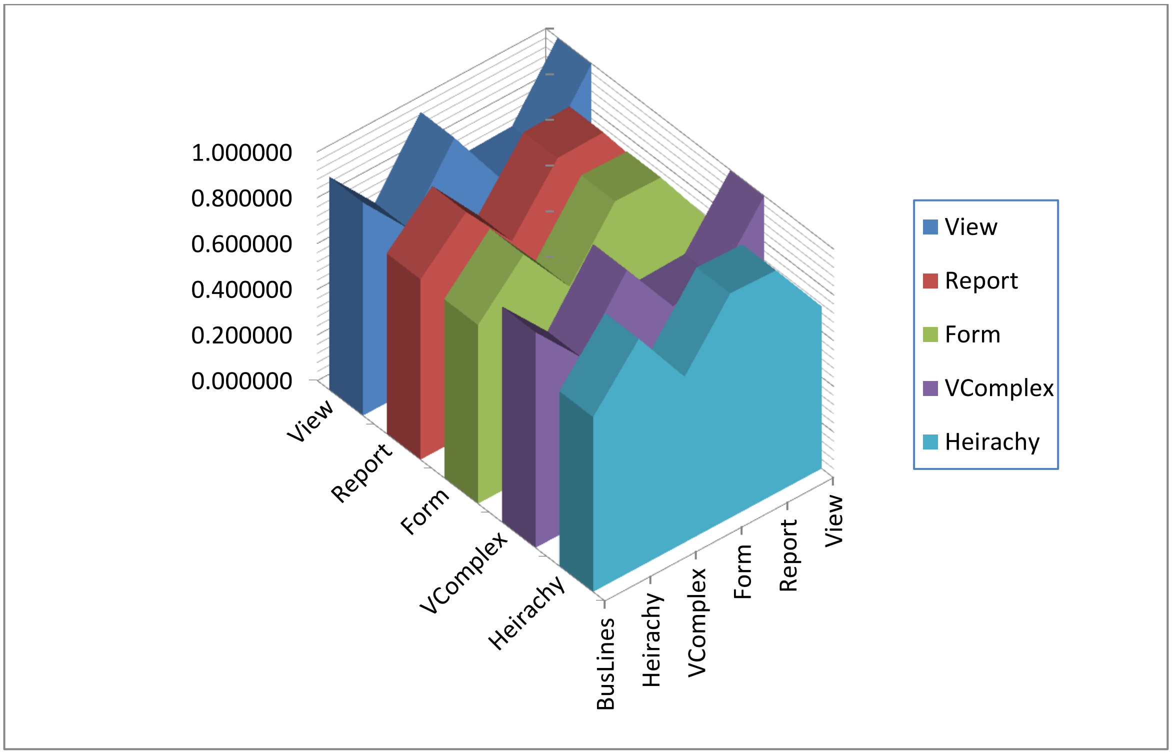


Figure 5-18 Correlation Relationship between Five Ranking Points

An analysis of Iteration 2 showed that the mechanics of the approach were appropriate and so additional ranking points were introduced. Using method described by Dunne et al., 2009, p853 (See 2.7.2, Business Weighting, p29) the reports are classified by usage based on the four hierarchal classifications of employees.

5.3.3 *Iteration 3 - Conclusion*

This iteration introduced a report usage value weighting factor that biased contestable data element ranking based on the end-user's position on the organisational hierarchical structure. The North Wind Trader's database had four levels of 'ReportsTo".

In an operational database, there may be more levels of employees in the hierarchical structure. This would be the subject of future research.

5.4 Iteration 4 – Six Ranking Points added and Recalibrated

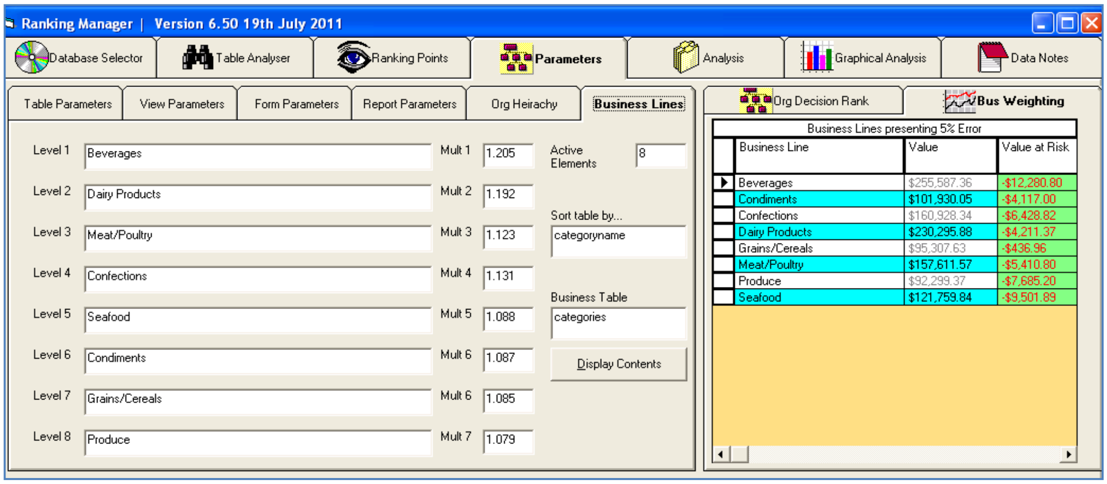
5.4.1 *Observation*

A new ranking point was introduced that graded contestable data elements based on their financial contribution to the product lines as described in (2.7.2, *Business Weighting*, p29)

The weighting was trialed using 0 (no effect), reverse weighting (lower correlation) and the final weighting shown below:

1. A facility was designed and developed that allowed additional ranking points to be included that better reflected the decision usage of various contestable data elements as well as the financial value (at risk) represented by the various contestable data elements.

   This facility is illustrated below showing the sales categories with total value and the amount-at-risk due to the seeded data quality errors. Note that the figure does not represent the value (or costs) of the data quality error, but does illustrate the variable propagation of data quality errors (5% in this instance) that present financial risks against different product lines. More about this in (6, *Experiment* , p76)



Level 1	Beverages	Mult 1	1.205
Level 2	Dairy Products	Mult 2	1.192
Level 3	Meat/Poultry	Mult 3	1.123
Level 4	Confections	Mult 4	1.131
Level 5	Seafood	Mult 5	1.088
Level 6	Condiments	Mult 6	1.087
Level 7	Grains/Cereals	Mult 6	1.085
Level 8	Produce	Mult 7	1.079

Active Elements: 8

Sort table by...: categoryname

Business Table: categories

Business Lines presenting 5% Error

Business Line	Value	Value at Risk
Beverages	$255,587.36	-$12,280.80
Condiments	$101,930.05	-$4,117.00
Confections	$160,928.34	-$6,428.82
Dairy Products	$230,295.88	-$4,211.37
Grains/Cereals	$95,307.63	-$436.96
Meat/Poultry	$157,611.57	-$5,410.80
Produce	$92,299.37	-$7,685.20
Seafood	$121,759.84	-$9,501.89

5.4.2 *ANALYSIS*

The results were used to bias contestable data elements that participated in the higher ranked views.

Note that many contestable data elements participated in many views at the different ranked levels, and so acquired a higher rank rating at several ranking points.

5.4.3 *TEST*

The correlation Grid was regenerated and the results observed:

Table 5-6 TEST: Correlation of Coefficient for N Elements for Six Ranking Points

	View	Report	Form	View Ranking	Hierarchy	Business Lines
View	1.000000	0.720802	0.715360	0.999125	0.709555	0.931988
Report	0.720802	1.000000	0.995633	0.740439	0.975565	0.788972
Form	0.715360	0.995633	1.000000	0.734334	0.983197	0.782385
View Ranking	0.999125	0.740438	0.734334	1.000000	0.725979	0.943741
Hierarchy	0.709555	0.975565	0.983197	0.725979	1.000000	0.765225
Business Lines	0.931988	0.788972	0.782385	0.943741	0.765225	1.000000

The correlation table above shows that the relationship between Views, Reports, Forms, and View Complexity, Hierarchical order and Business Lines is significant.

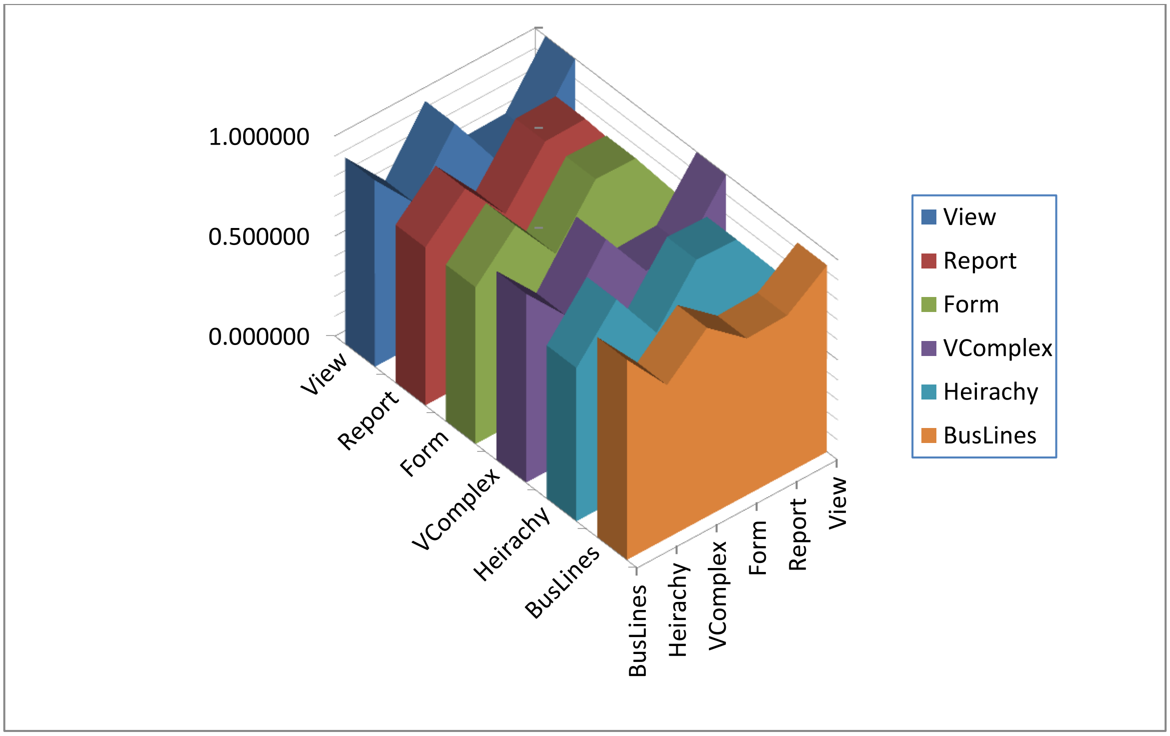
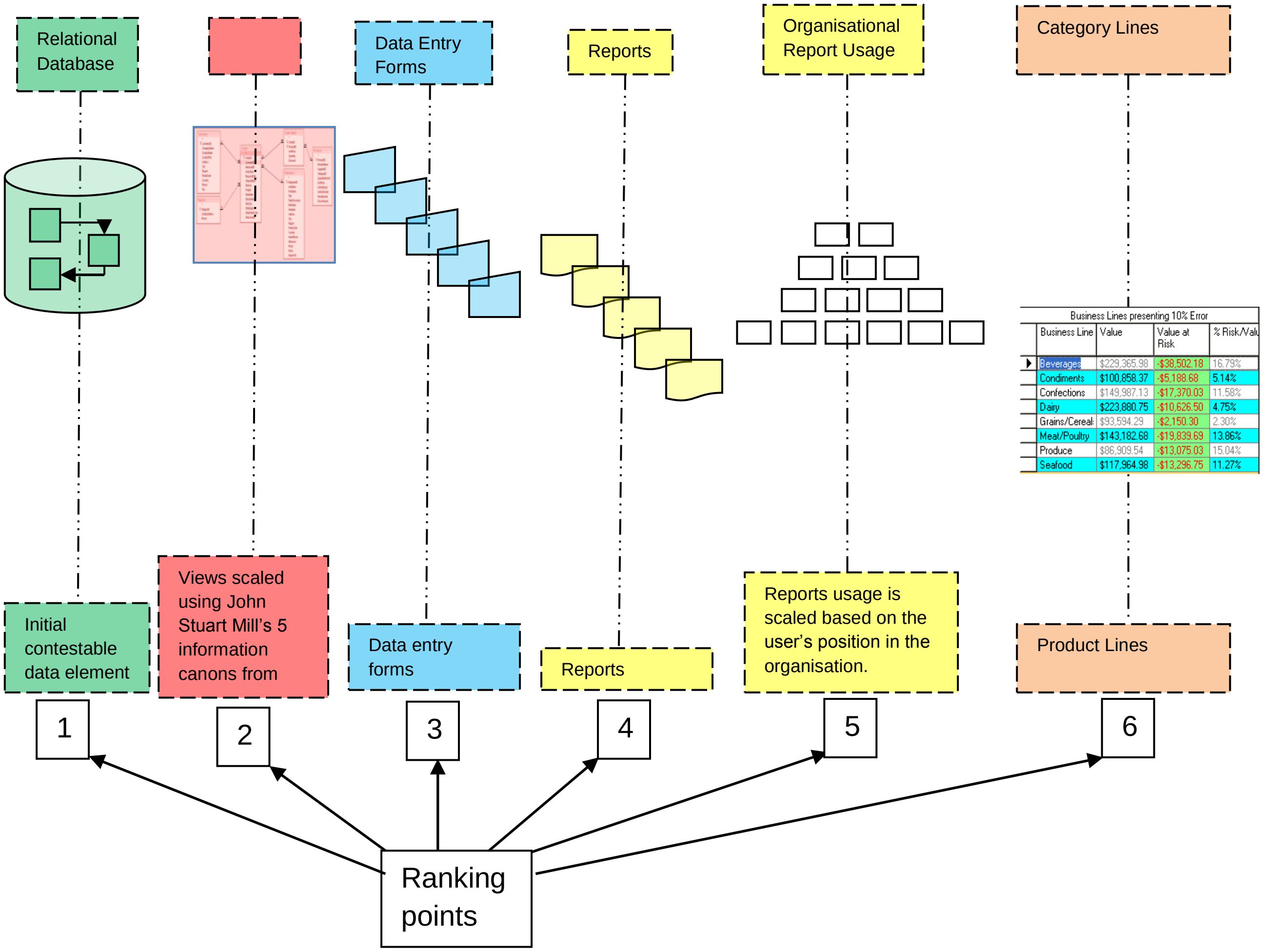


Figure 5-19 Correlation Relationship between Six Ranking Points

The correlation table has been graphed to show the relationships between the correlations of the ranking points. The commonality of patterns between (*Business Lines, View Complexity and Views*) as against (*Hierarchical Report Usage, Reports, Forms*) is noted but at this stage unexplained. This is the subject of further research.

5.4.4 *Iteration 4 - Conclusion*

This approach now offers a six ranking points set:



Relational Database
Data Entry Forms
Reports
Organisational Report Usage
Category Lines
Initial contestable data element
Views scaled using John Stuart Mill's 5 information canons from
Data entry forms
Reports
Reports usage is scaled based on the user's position in the organisation.
Product Lines
1
2
3
4
5
6
Ranking points

	Business Lines presenting 10% Error			
	Business Line	Value	Value at Risk	% Risk/Valu
▶	Beverages	$229,365.98	-$38,502.18	16.79%
	Condiments	$100,858.37	-$5,188.68	5.14%
	Confections	$149,987.13	-$17,370.03	11.58%
	Dairy	$223,880.75	-$10,626.50	4.75%
	Grains/Cereal:	$93,594.29	-$2,150.30	2.30%
	Meat/Poultry	$143,182.68	-$19,839.69	13.86%
	Produce	$86,909.54	-$13,075.03	15.04%
	Seafood	$117,964.98	-$13,296.75	11.27%

Figure 5-20 Ranking Measurement Points - 4th Iteration

5.5 Iteration 5 – Experimental Implementation and Trials

5.5.1 *Observation*

A facility was required to allow each contestable data element to be identified and tracked as it participated in views, data entry forms, and reports. The initial trial was to add a four character key to each contestable data element and use this key to reference a contestable data element table and to allow each data element to be tracked.

The initial plan was to modify the contents of each contestable data element be altered by concatenating a key reference.

This iteration reflected a major design change that formed the basis of the current ranking tool state. A routine was devised that identified and examined each row in each table in the database. As each contestable data element was discovered, a new row entry was created in a data property table with columns ready for population as each ranking point was measured. This design also included a facility that presented a two-way linked-list to allow the property table to link to its source row and column in each table as well as linking from each row column interest to the property table.

The final iteration of the property table was designed to allow the random ‘seeding’ of data quality flags to be displayed along with the scalar values for each contestable data element. This then allowed experiments to be conducted to test different techniques for detection data quality errors. More about this in (6, *Experiment* , p76)

Table 5-7 Contestable Data Element Property Extensions

Column Name	**Data Type**
Rank_Database	Text
Rank_TableName	Text
Rank_Column	Text
Rank_PrimaryID	Text
Rank_PrimaryField	Text
Rank_RowNumber	Single
Rank_FieldType	Text
Rank_FieldContent	Text
Rank_Flagged	Yes/No
Rank_Table	**Single**
Rank_View	**Single**
Rank_Report	**Single**
Rank_Form	**Single**
Rank_ViewCmplx	**Single**
Rank_Heirachy	**Single**
Rank_BusLine	**Single**

5.5.2 *ANALYSE*

This section describes some of the issues and solutions encountered during the development of the artifact.

1. Initially, the plan was to introduce actual data errors into the database.

   This approach introduced a number of operational errors. The North Wind database comprises data that contains many character modifiers (grave, accent, umlaut and so on) the ASCII range available in contestable data element strings became too small. Trials were also conducted to introduce data quality errors (as a flag) randomly across the entire database. Upon inspection, there were two key issues noted. Strings could be flagged as ‘error on’, but numeric fields were more problematic. Making a number negative (signing the lead bit) would have been valid with the North Wind Traders database (there were no negative numeric fields), it would have constrained further use of the artifact.

2. Establishing Boolean flags against each contestable data element addressed the issues noted in iteration 1. The design also allows all contestable data elements to be mapped in a continuous string (wrapped each 120 characters per line) that allowed a distribution of the ‘errors’ to be displayed.

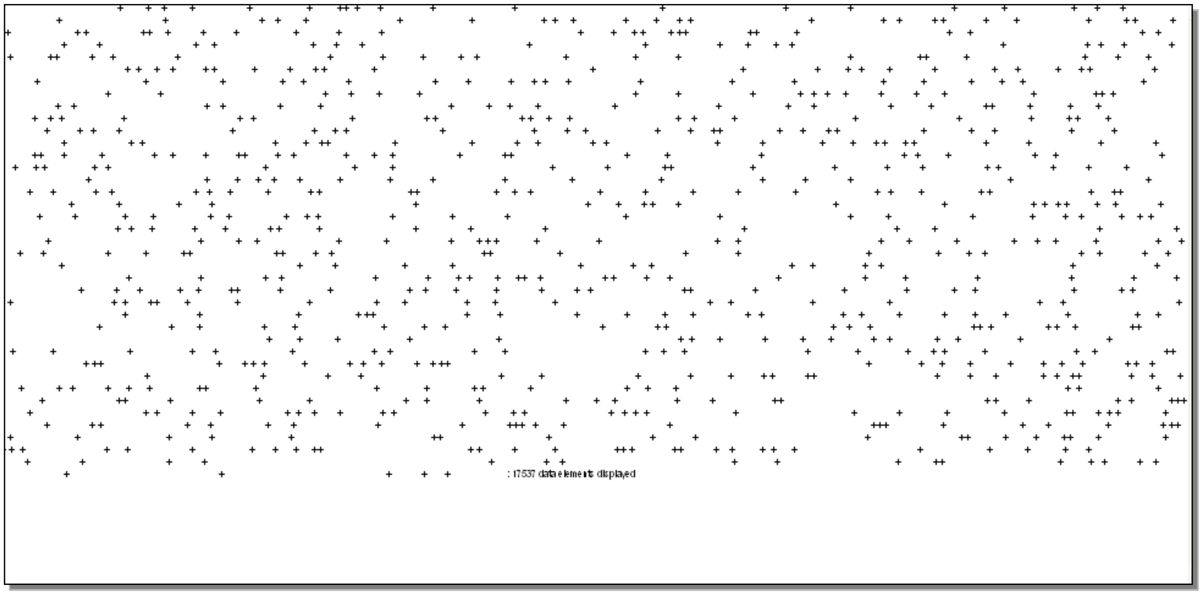
3. As the analysis progressed that the six accumulation points were quite different topics and so the accumulation characteristics suggested that a single vector ranking approach would not exploit these characteristics effectively and therefore may not offer a balanced scalar value for each contestable data element.

   To that end, six ranking points were created. Each contestable data element participating in each measurement point then has its scoring value incremented.

   A graphics display was created and some numeric tests to determine random distribution was developed as a quality measure. The rules used were that if any random place (between 1 and 17537) was already occupied with an error flag, then the cycle was repeated until a non-set flag place was detected.

   Once the randomization was completed, the random flags were copied to the Rank_table file and the relevant contestable data elements were flagged as true.

   For example, the 17537 contestable data elements are represented in a single string that has ‘+’ for flagged and “ “ for not flagged. This string is displayed from left to right, whilst wrapping at the right boundary to show the random distribution of flagged data elements.



: 17537 data elements displayed

Figure 5-21 Random Quality Error Seeding Display at 10% Errors

The 17537 represents the number of contestable data elements in the North Wind Traders database. This facility allows observation of the 'randomness' of the distribution.

5.5.3 *TEST*

Different percentage error rates were trialed. 5% is prominent in literature review (See 2.5, Cost and Value of Data Quality, p18) and so the experiments have been set at 1%, 5% and 10% to offer bracketing of the expected common data quality error rate. For detailed analysis, see the attached CD

As the percentage rates became less than 0.1%, division errors occurred as many of the calculations showing aggregated error rates failed as the rate was too low to have generated an error data. This error rate is considered insignificant.

As the percentage error rates approached 100% the time taken to complete the seeding became significant. This is because the random seeding generator checks to see if a contestable data element has already been seeded. If the system strikes a seeded contestable data element, then is regenerates a new random number and attempts to seed again. As the seed approach 100% the reseeding 'strikes' become substantial and impractical. Data quality error simulation beyond 20% would be of little investigative value.

5.5.4 *ITERATION 5 - CONCLUSION*

The ranking points of this final iteration are:

[Rank_View]
This ranking point is applied to all contestable data elements as each is detected in each view. The scalar value has been determined by applying a common scalar value each time a contestable data element is detected in each row in each view.

[Rank_Report]
This ranking point is uniformly applied to all contestable data elements each time a contestable data element is detected in each row in each report.

[Rank_Form]
This ranking point is uniformly applied to all contestable data elements using a CSV. This value represents the base scalar value for all elements.

[Rank_ViewComplexity]
This ranking point is applied to all contestable data elements participating in one or more views based on their ranking using a view ranking mechanism. Note that a contestable data element may participate in many views and accumulate higher ranking measures as a result.

[Rank_ProductLine]

This ranking point is applied to all contestable data elements participating in one or more product lines based on their ranking using a proportional product line value ranking. Note that a contestable data element may participate in many product lines and accumulate higher ranking measures as a result.

6 EXPERIMENT FINDINGS

Chapter Introduction

This chapter describes the tests that were conducted against the target database showing the difference between the current un-ranked detect-and-treat approach, ranked detect-and-treat and, as a control, a ‘perfect’ world where the data errors are known and ranked in highest to lowest order.

The database was seeded with various percentages of data quality errors to simulate data quality errors.

Each contestable data element is assigned a base scalar value of one. These data elements accrue a scalar value based on the number of times they are used, the purpose of the usage, the hierarchical ranking of the data user (via reports) and the value of the product lines to which they contribute.

The product lines used to demonstrate the " Product value at risk" figures could, in other databases be other expenditure or income streams, performance indicators, production areas in an organisation, accounts clusters or supply chain targets.(See 2.4, Data Quality in Supply Chains, p14)

6.1 THE CONDUCT OF THE EXPERIMENT

This section describes and illustrates the differences that varying percentages of simulated data quality errors and carrying ‘correction’ rates against each simulation. Note that these figures are specific to the subject database and the corporate structure implied in the North Wind Trader’s database. Different databases in a different organisational structure may present different results.

Given that the percentage of actual data quality errors appears to be around 5%, this was initially the focal point for the experimentation. (See 2.1, Data Quality and Decision making, p7)

Varying the percentage of simulated errors, it became apparent that the variations between percentages affected the benefits of ranking the contestable data element in a database.

In addition a simulated clearing of the data quality errors was designed so that for a common effort each of the three approaches was trialed for each percentage.

The three processes trialed in each experiment were:

1. Unknown quality error points using unranked data;

2. Unknown quality error points using ranked data; and

3. Known quality errors points using ranked data.

The difference in effectiveness between Process 1 and Process 2 represents the advantages associated with data ranking using the ranked approach described in this thesis.

Process 3 is a control mechanism where the data quality errors are known and ranking in order of importance.

There are four sets of experiments detailed with 1%, 5%, 10% and 15% simulated data quality errors for each of the processes showing the approach and the outcome for each.

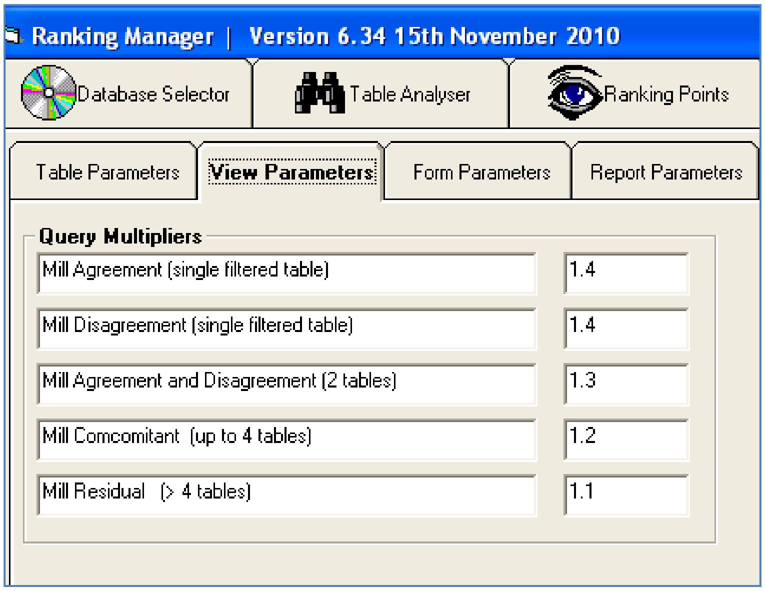
These ranges were selected to bracket the effectiveness of the ranked data element approach with the highest effective point being between 4% and 12%. This depends upon the database and the organisation's usage profile. Although the advantages of the ranked data approach are evident across the ranges, 1% error rate is likely to be a low error rate that would be acceptable to many organisations; 15% may be considered too high in many commercial applications.

These experiments show how the ranking method improves data quality as a scalar value clearance when compared to error detection using unranked data elements.

Each of these experiments shows the conduct and outcomes of the four different data quality error rates.

6.1.1 EXPERIMENTAL CONSTANTS

The constants are shown here can be varied for other experiments, but for these experiments are constant.



Query Multipliers

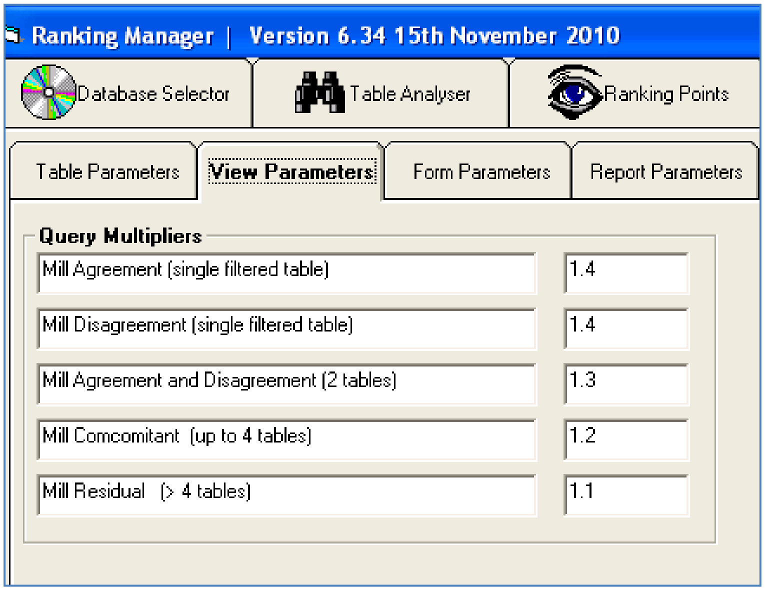
Mill Agreement (single filtered table)	1.4
Mill Disagreement (single filtered table)	1.4
Mill Agreement and Disagreement (2 tables)	1.3
Mill Comcomitant (up to 4 tables)	1.2
Mill Residual (> 4 tables)	1.1

Figure 6-1 Ranking View Parameters (Data Facets)

Figure 6-1 Ranking View Parameters (Data Facets) shows the classifications of the queries used in the database. The rationale for this approach is explained in 2.7, Ranking Approaches, p24.

Table 6-1 Tables, Views, Forms and Reports

	Tables	Views	Forms	Reports
Contestable Data Elements	17397	182650	403044	719782



Query Multipliers

Mill Agreement (single filtered table)	1.4
Mill Disagreement (single filtered table)	1.4
Mill Agreement and Disagreement (2 tables)	1.3
Mill Comcomitant (up to 4 tables)	1.2
Mill Residual (> 4 tables)	1.1

Figure 6-1 Ranking View Parameters (Data Facets)

Figure 6-1 Ranking View Parameters (Data Facets) shows *the classifications of the queries* used in the database. The rationale for this approach is explained in 2.7, Ranking Approaches, p24.
 shows the 'multiplier effect' of data elements as they are tracked and used in one or more views, forms and reports.

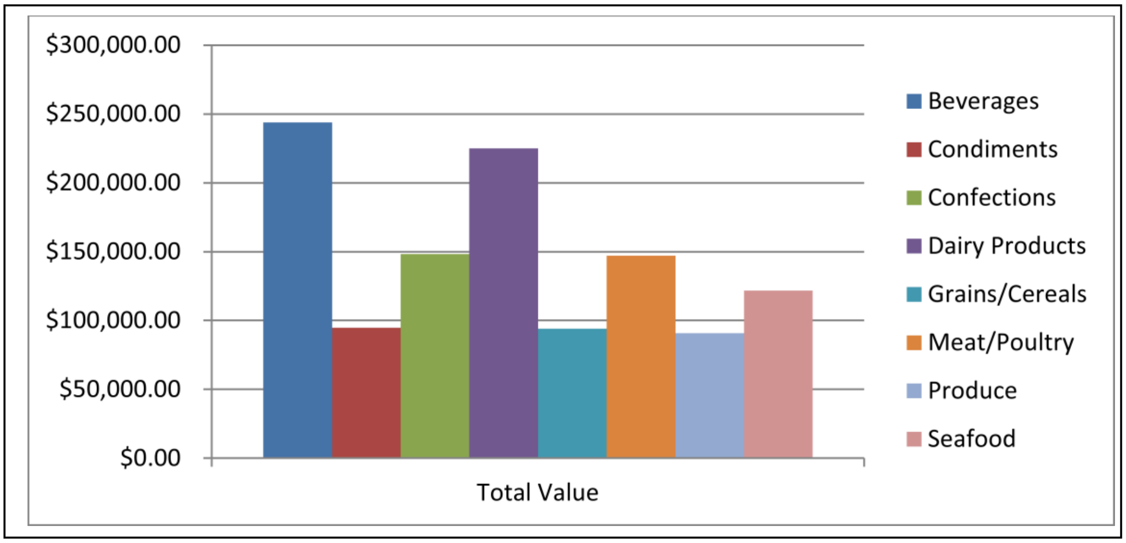
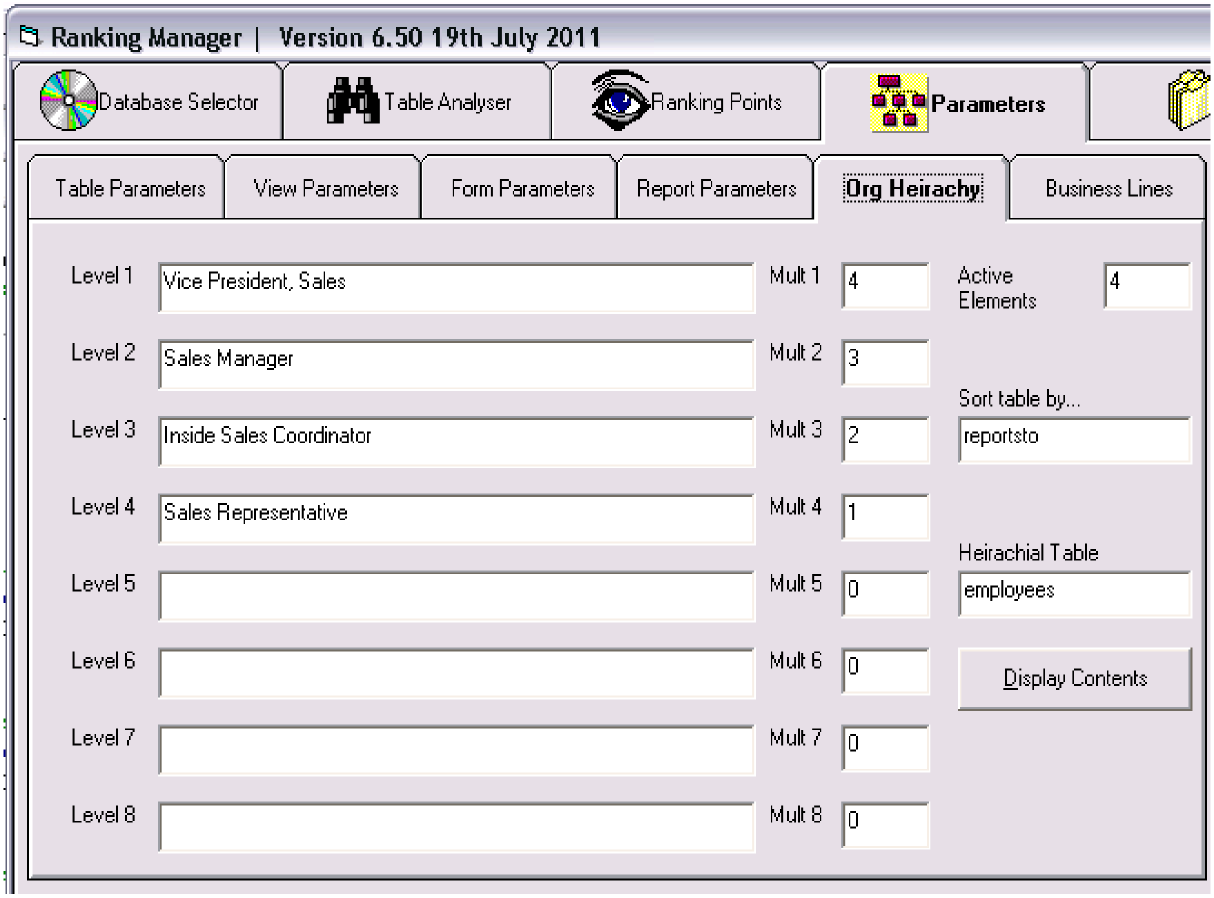


Figure 6-2 Product Lines by total value

Figure 6-2 Product Lines by total value graphically shows the total values of the product lines. This type of display is typical for many organisations, but does not indicate the possible percentage of data errors that might be present in terms of errors of some value indication.



Level 1	Vice President, Sales	Mult 1	4
Level 2	Sales Manager	Mult 2	3
Level 3	Inside Sales Coordinator	Mult 3	2
Level 4	Sales Representative	Mult 4	1
Level 5		Mult 5	0
Level 6		Mult 6	0
Level 7		Mult 7	0
Level 8		Mult 8	0

Active Elements: 4

Sort table by... reportsto

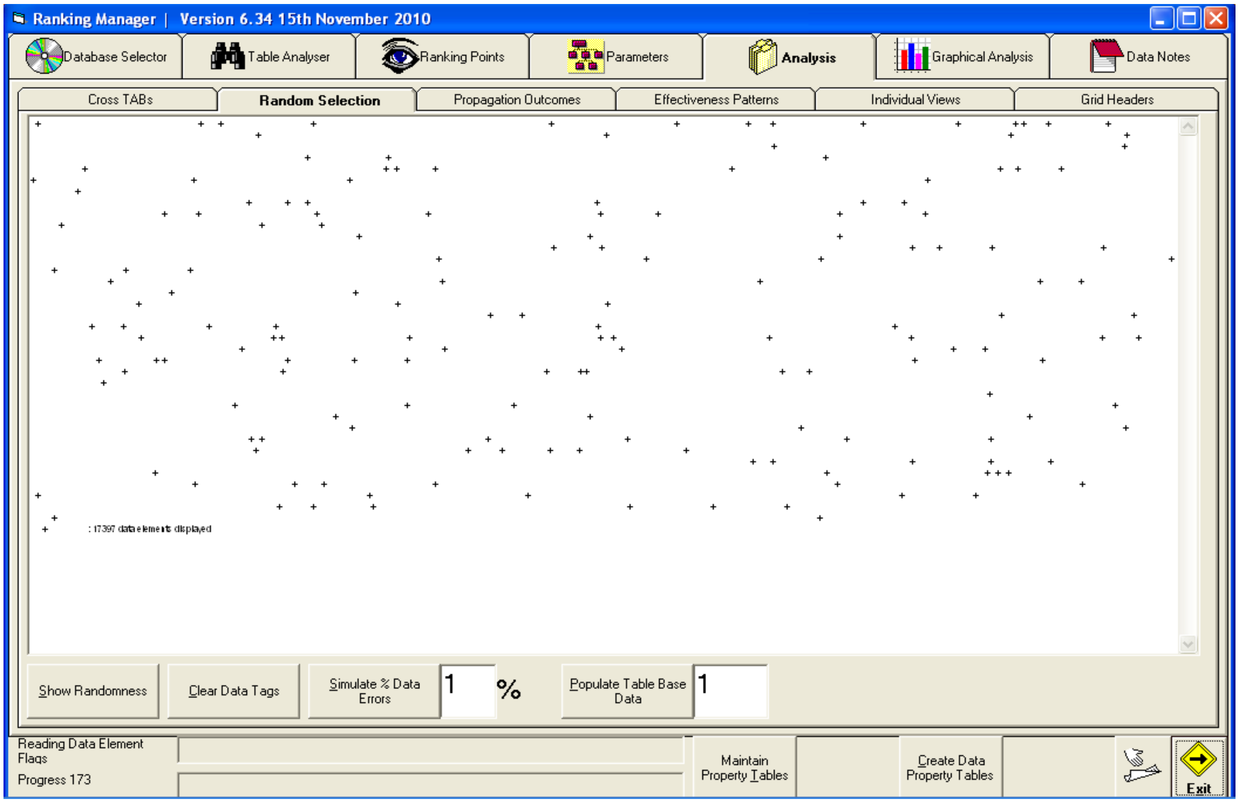
Heirachial Table: employees

Figure 6-3 'Reports to' report weighting for all experiments

Figure 6-3 'Reports to' report weighting for all experiments shows the weighting assigned to different organisational ranking structures. These weightings are used to further weight the data elements that contribute to reports used by different users in the organisation. There are data elements that contribute to more than one report and so they weight higher than data elements that only contribute to lower ranked or less reports.

6.2 EXPERIMENTS USING 1% SIMULATED ERRORS

A typical 1% simulated data error rate offers the following distribution of simulated errors across all contestable data elements:



: 17397 data elements displayed

Simulate % Data Errors 1 %

Populate Table Base Data 1

Reading Data Element Flags

Progress 173

Figure 6-4 1% Simulated Error Random Display

Figure 6-4 1% Simulated Error Random Display shows the simulated data error propagation. This display shows that the data quality error simulation is randomly distributed across all contestable data elements.

Table 6-2 1% data quality error propagation in Tables, Views, Forms and Reports

	Tables	**Views**	**Forms**	**Reports**
Contestable Data Elements	17397	182650	403044	719782
Flagged Elements	173	1074	332	677

Table 6-2 1% data quality error propagation in Tables, Views, Forms and Reports shows that the data quality errors propagate at different levels across views, forms and reports. This propagation ratio appears proportionately similar across all tested percentage data quality error simulations against the North Wind Trader's database.

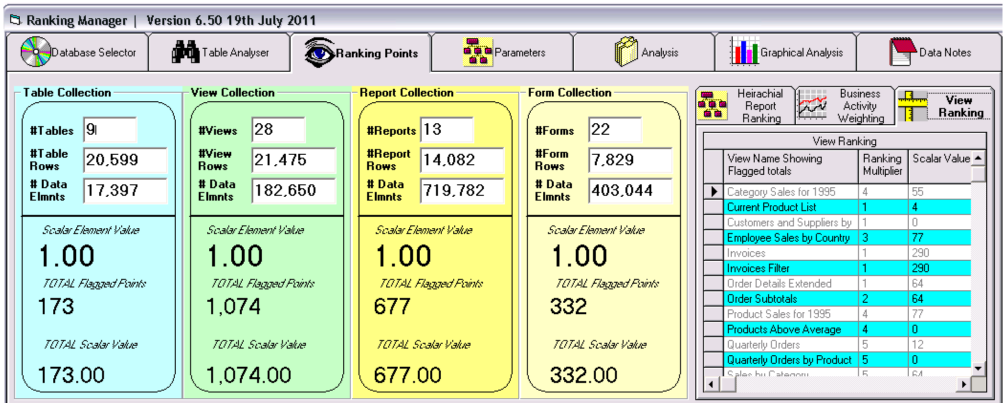


Table Collection

#Tables 9
#Table Rows 20,599
Data Elmnts 17,397
Scalar Element Value 1.00
TOTAL Flagged Points 173
TOTAL Scalar Value 173.00

View Collection

#Views 28
#View Rows 21,475
Data Elmnts 182,650
Scalar Element Value 1.00
TOTAL Flagged Points 1,074
TOTAL Scalar Value 1,074.00

Report Collection

#Reports 13
#Report Rows 14,082
Data Elmnts 719,782
Scalar Element Value 1.00
TOTAL Flagged Points 677
TOTAL Scalar Value 677.00

Form Collection

#Forms 22
#Form Rows 7,829
Data Elmnts 403,044
Scalar Element Value 1.00
TOTAL Flagged Points 332
TOTAL Scalar Value 332.00

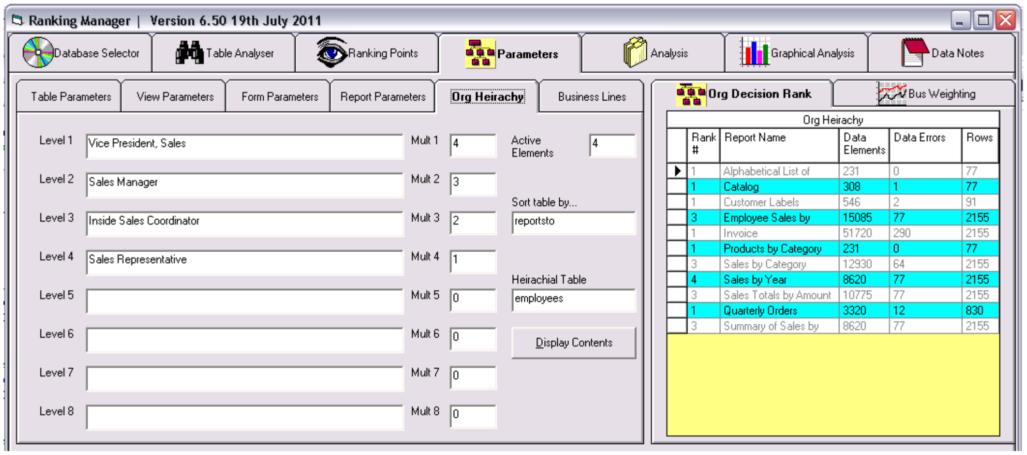
View Ranking

View Name Showing Flagged totals	Ranking Multiplier	Scalar Value
Category Sales for 1995	4	55
Current Product List	1	4
Customers and Suppliers by	1	0
Employee Sales by Country	3	77
Invoices	1	290
Invoices Filter	1	290
Order Details Extended	1	64
Order Subtotals	2	64
Product Sales for 1995	4	77
Products Above Average	4	0
Quarterly Orders	5	12
Quarterly Orders by Product	5	0
Sales by Category	5	64

Figure 6-5 1% Data Quality Error Propagation Rate (Flagged Points)

Figure 6-5 1% Data Quality Error Propagation Rate shows the baseline figures that represent the database with a 1% seeded error rate.

Report Ranking by usage where the usage numbers are represented by the hierarchical order of the report users:



Level 1 Vice President, Sales — Mult 1 4
Level 2 Sales Manager — Mult 2 3
Level 3 Inside Sales Coordinator — Mult 3 2
Level 4 Sales Representative — Mult 4 1
Level 5 — Mult 5 0
Level 6 — Mult 6 0
Level 7 — Mult 7 0
Level 8 — Mult 8 0

Active Elements 4

Sort table by... reportsto

Heirachial Table employees

Org Heirachy

Rank #	Report Name	Data Elements	Data Errors	Rows
1	Alphabetical List of	231	0	77
1	Catalog	308	1	77
1	Customer Labels	546	2	91
3	Employee Sales by	15085	77	2155
1	Invoice	51720	290	2155
1	Products by Category	231	0	77
3	Sales by Category	12930	64	2155
4	Sales by Year	8620	77	2155
3	Sales Totals by Amount	10775	77	2155
1	Quarterly Orders	3320	12	830
3	Summary of Sales by	8620	77	2155

Figure 6-6 1% Data Quality Error by Report Hierarchical Usage

Figure 6-6 1% Data Quality Error by Report Hierarchical Usage shows that there are low error rates present in reports. Some reports (at 1%) do not have any flagged data elements. This is due to low percentage error rates where some reports will exhibit zero errors for some random trials.

Table 6-3 Product value at risk for 1% error simulation

Sales Category	Total Value	Value at Risk	% Value at Risk
Beverages	$266,500.01	-$1,368.15	0.51%
Condiments	$105,985.05	-$ 62.00	0.06%
Confections	$164,804.79	-$2,552.36	1.55%
Dairy Products	$232,261.45	-$2,245.80	0.97%
Grains/Cereals	$ 94,889.59	-$ 855.00	0.90%
Meat/Poultry	$162,520.77	-$ 501.60	0.31%
Produce	$ 99,210.57	-$ 774.00	0.78%
Seafood	$127,834.13	-$3,427.60	2.68%

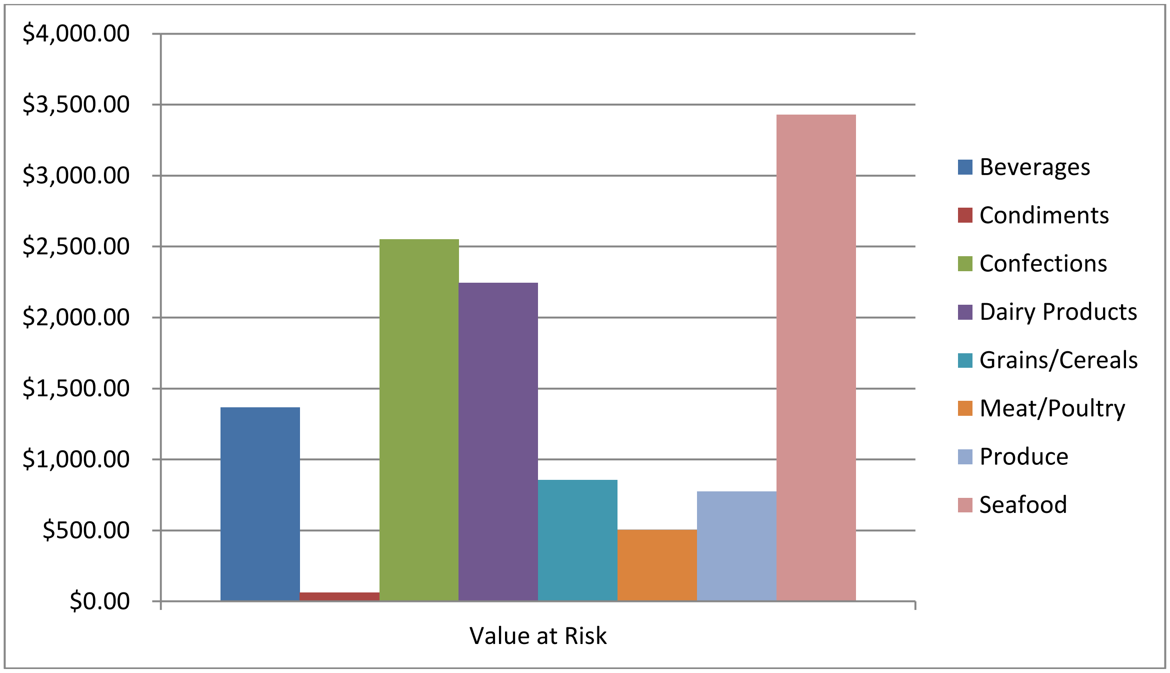


Figure 6-7 Product Line value-at-risk for 1% error simulation

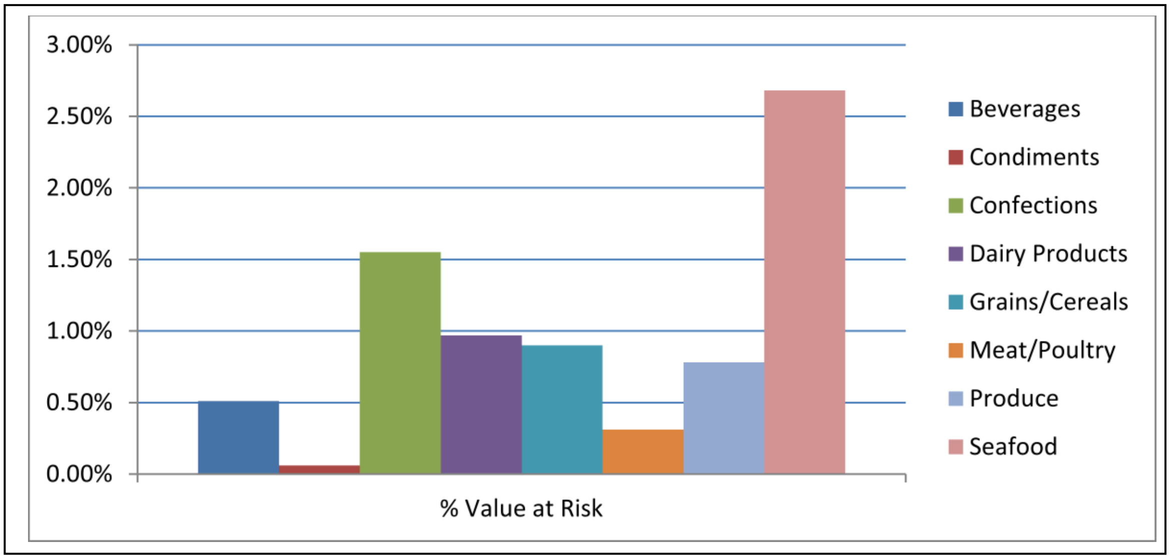


Figure 6-8 Product Line percentage value risk for 1% data quality error simulation

In similar fashion to the report error rates (see *Figure 6-6 1% Data Quality Error by Report Hierarchical Usage*), the product line 'value at risk' rates are expectedly low.

The ‘Value at Risk’ is a representation of the derived values where at least one of the data elements in each formula contains a flagged data element.

The variance in that different business lines do not reflect the 1% error rate due to different volumes and price per unit variations.

6.3 Experiment Findings at 1%

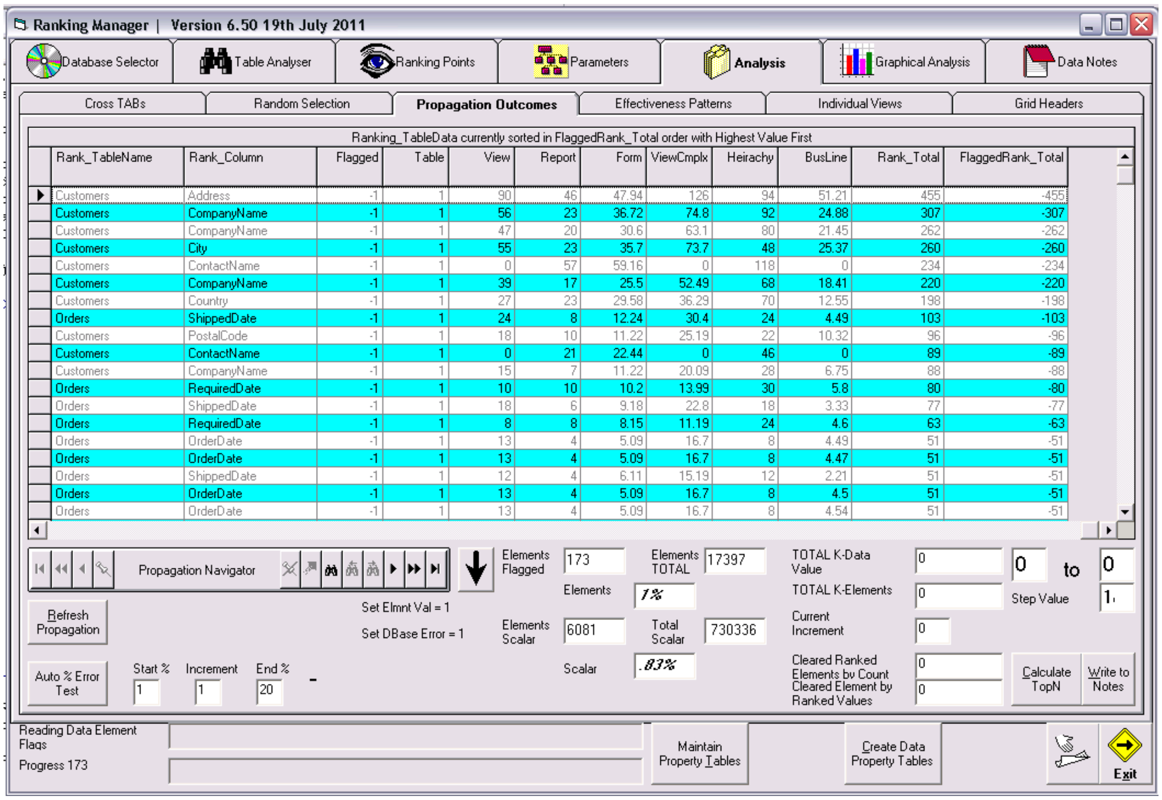
A 1% error rate in the North Wind Traders database would probably not present a major data quality issue for most organisations. This assumption matches much of the literature research that shows organisations reporting data quality issues with quality error rates around 5%. (*See Data Quality Error Rates p23*)

If an organisation's data was sensitive to any errors (such as a medical diagnosis data repository), then 1% error rate may be unacceptable.

The graph below illustrates the comparative correction rate for the three different approaches to detecting data quality errors.

Log entry:
Database: Northwind.mdb. Contains 17397 contestable data elements. These elements have been tagged representing a 1% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8314

The ranked table and data element flags are shown in *Figure 6-9 1% Error Rate Simulation* :



Ranking_TableData currently sorted in FlaggedRank_Total order with Highest Value First

Rank_TableName	Rank_Column	Flagged	Table	View	Report	Form	ViewCmplx	Heirachy	BusLine	Rank_Total	FlaggedRank_Total
Customers	Address	-1	1	90	46	47.94	126	94	51.21	455	-455
Customers	CompanyName	-1	1	56	23	36.72	74.8	92	24.88	307	-307
Customers	CompanyName	-1	1	47	20	30.6	63.1	80	21.45	262	-262
Customers	City	-1	1	55	23	35.7	73.7	48	25.37	260	-260
Customers	ContactName	-1	1	0	57	59.16	0	118	0	234	-234
Customers	CompanyName	-1	1	39	17	25.5	52.49	68	18.41	220	-220
Customers	Country	-1	1	27	23	29.58	36.29	70	12.55	198	-198
Orders	ShippedDate	-1	1	24	8	12.24	30.4	24	4.49	103	-103
Customers	PostalCode	-1	1	18	10	11.22	25.19	22	10.32	96	-96
Customers	ContactName	-1	1	0	21	22.44	0	46	0	89	-89
Customers	CompanyName	-1	1	15	7	11.22	20.09	28	6.75	88	-88
Orders	RequiredDate	-1	1	10	10	10.2	13.99	30	5.8	80	-80
Orders	ShippedDate	-1	1	18	6	9.18	22.8	18	3.33	77	-77
Orders	RequiredDate	-1	1	8	8	8.15	11.19	24	4.6	63	-63
Orders	OrderDate	-1	1	13	4	5.09	16.7	8	4.49	51	-51
Orders	OrderDate	-1	1	13	4	5.09	16.7	8	4.47	51	-51
Orders	ShippedDate	-1	1	12	4	6.11	15.19	12	2.21	51	-51
Orders	OrderDate	-1	1	13	4	5.09	16.7	8	4.5	51	-51
Orders	OrderDate	-1	1	13	4	5.09	16.7	8	4.54	51	-51

Elements Flagged 173
Elements TOTAL 17397
Elements 1%
Set Elmnt Val = 1
Set DBase Error = 1
Elements Scalar 6081
Total Scalar 730336
Scalar .83%
TOTAL K-Data Value 0
TOTAL K-Elements 0
Current Increment 0
Cleared Ranked Elements by Count 0
Cleared Element by Ranked Values 0
0 to 0
Step Value 1.
Start % 1
Increment 1
End % 20
Reading Data Element Flags
Progress 173

Figure 6-9 1% Error Rate Simulation

Figure 6-9 1% Error Rate Simulation shows the ranking outcomes for each of the ranking points for each contestable data element sorted into rank total from highest to the lowest. 173 data elements from 17397 have been flagged as having data quality errors representing 1% error rate.

Table 6-4 1% error rate. Clearance rate in Unranked *and* Ranked order

Examination Effort		Unranked Data		Ranked Data		Effectiveness
Data Elements Examined %	Data Elements Examined	Cumulative Scalar Value	Cumulative Unit Count	Cumulative Scalar Value	Cumulative Unit Count	Ranked Compared *to* Unranked
1%	173	**4592**	5	**173**	1	-2554.3
2%	347	**5583**	12	**762**	2	-632.7
3%	521	**6138**	21	**1936**	7	-217.0
4%	695	**7979**	29	**1936**	7	-312.1
5%	869	**6589**	35	**2135**	9	-208.6

Table 6-4 1% error rate. Clearance rate in Unranked and Ranked order shows the constant remediation effort and then the effect against unranked and ranked data in terms of unit clearance rates as well as scalar clearance rates.

In this table the ranked cleared unit count is less than the unranked unit count as well as the cumulative scalar values. This outcome shows that ranking data elements offers low or negative value with very low unit data quality rates.

This table also demonstrates the high negative count of data quality examination when compared to the yield from the examination (Unranked data 173-5=168, Ranked data 173-1=172).

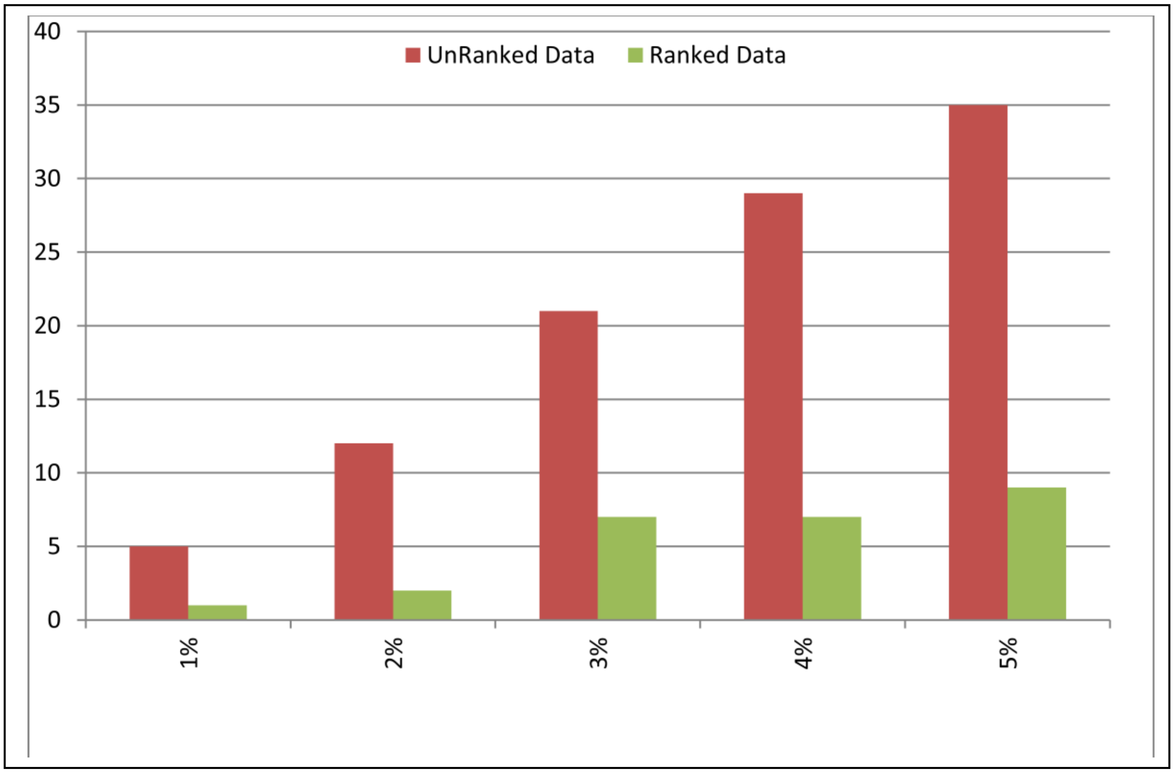


Figure 6-10 1% Unranked Vs Ranked Data - Data Elements Cleared

Figure 6-10 1% Unranked Vs Ranked Data - Data Elements Cleared shows that the difference between ranked and unranked data clearance rates is minor and, in practice, may not offer any advantages at all. The illustrated difference is a function of random seeding. In both examples of ranked and unranked the differences are very small.

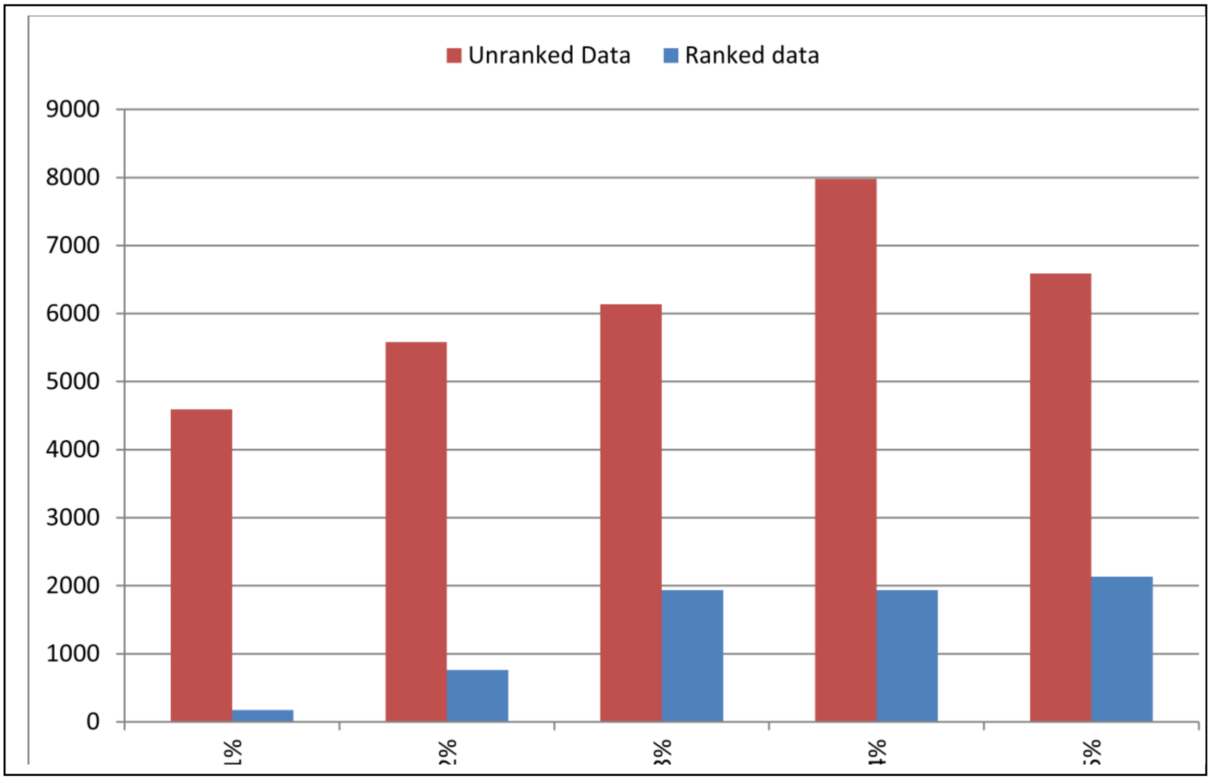


Figure 6-11 1% to 5% Unranked Vs Ranked Data - Scalar Values Detected

Similarly to the unit count detection rate, this ranking detection rate is actually negative when compared to the unranked order. This is caused by the very low success rate caused by the 1% random error seeding. Using a low error seeding rate means that the first 1% of detection is likely to represent much of the scalar value of the data in error.

The contestable data elements are ‘cleared’ and the various proportions of the data errors in increments of 1% to 5% to see the difference between random detection, ranked detection and (the control) ranked detection with data errors known.

The random detection approach yields very low data quality error ‘hit rates’ and a small amount of scalar ranked data.

This table shows slightly different detection rates (the data has been ordered by ranking this time instead of some random order). As expected, the number of data units detected with quality errors is not improved. In this example it turns out to be worse that random selection.

6.3.1 *KNOWN QUALITY ERRORS POINTS USING RANKED DATA*

This is the control component where the data is ranked in scalar order and in flagged order.

Table 6-5 CONTROL at 1% error. Clearance rate in Ranked and Error flag order

% Data Elements Examined	Data elements checked	Scalar Value	Flagged Data Units Detected
1%	173	6081	173
2%	347	6081	173
3%	521	6081	173
4%	695	6081	173
5%	869	6081	173

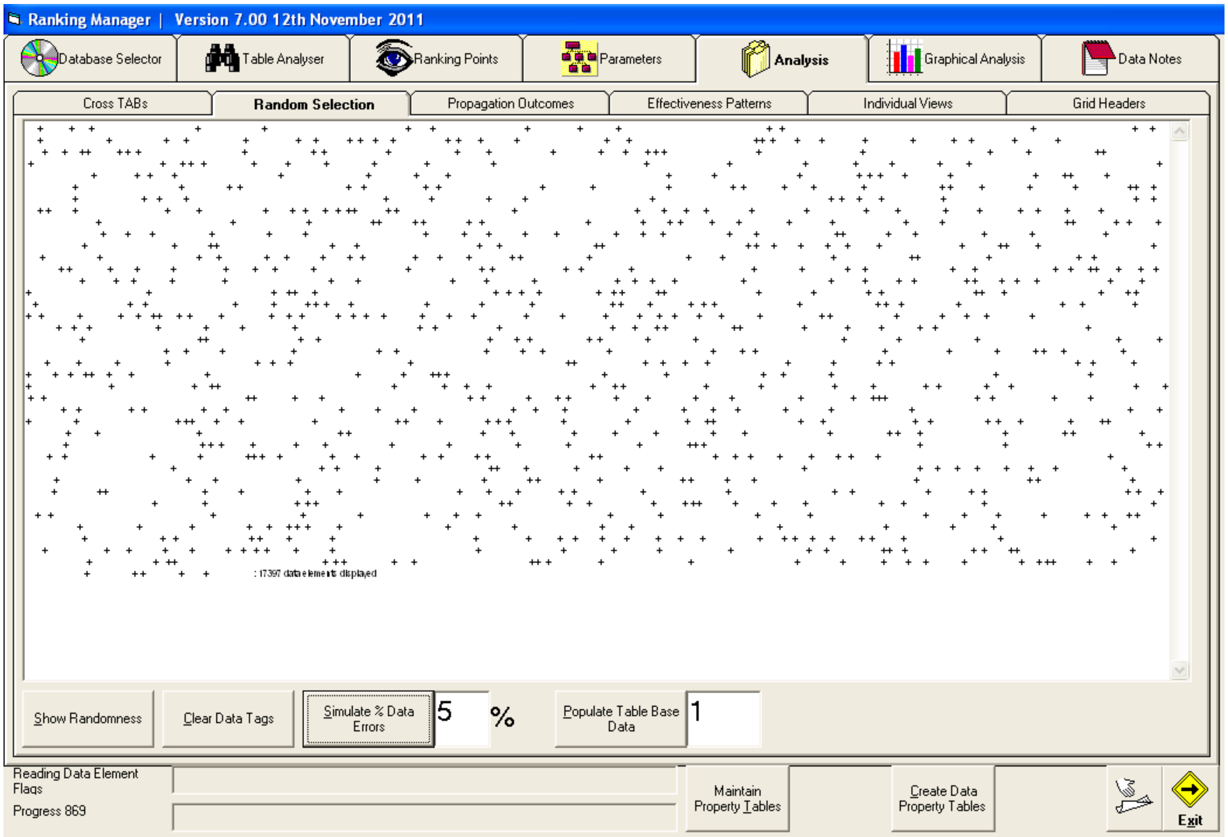
The control (being in data error and ranked order) clears all data errors in the first 1% of investigation.

The scalar value of the detected data element also roughly matches the random selection test. This demonstrates that with 1% data quality errors, there is little benefit using ranked data when compared to using unranked data.

In either case the effort and costs detecting error with this level of data quality errors is unlikely to be considered a valid expense by most organisations.

6.4 Experiments Using 5% Simulated Errors

A 5% error rate in the North Wind Traders database is considered typical of many organisation's data quality levels. This outcome matches much of the literature research where companies report data quality issues tended to report around 5% errors. (See 2.1, Data Quality and Decision making, p7 for examples).



Ranking Manager | Version 7.00 12th November 2011

: 17397 data elements displayed

Simulate % Data Errors 5 %

Populate Table Base Data 1

Reading Data Element Flags

Progress 869

Figure 6-12 5% Simulated Error Random Display

The simulation for 5% data quality errors yields 869 flagged data elements from 17397 in total.

Log Book entry:

Database: Northwind.mdb. Contains 17397 contestable data elements. These elements have been tagged representing a 5% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8678

Table 6-6 5% data quality error propagation in Tables, Views, Forms and Reports

	Tables	**Views**	**Forms**	**Reports**
Contestable Data Elements	17397	182650	403044	719782
Flagged Elements	869	7675	2692	5058

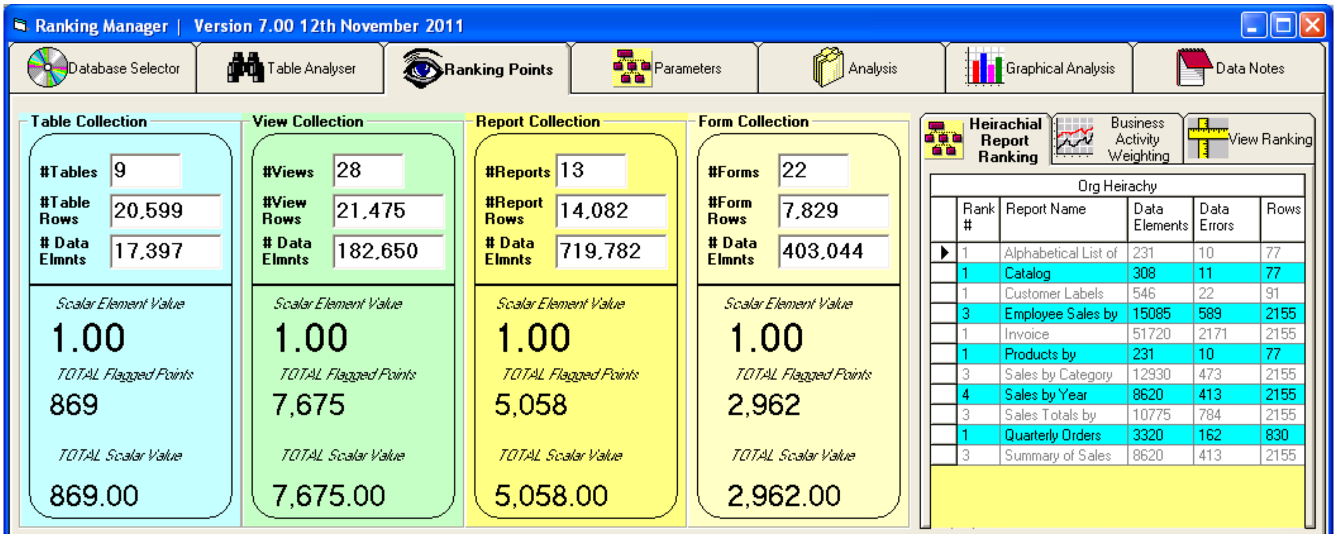


Table Collection

#Tables 9
#Table Rows 20,599
Data Elmnts 17,397
Scalar Element Value 1.00
TOTAL Flagged Points 869
TOTAL Scalar Value 869.00

View Collection

#Views 28
#View Rows 21,475
Data Elmnts 182,650
Scalar Element Value 1.00
TOTAL Flagged Points 7,675
TOTAL Scalar Value 7,675.00

Report Collection

#Reports 13
#Report Rows 14,082
Data Elmnts 719,782
Scalar Element Value 1.00
TOTAL Flagged Points 5,058
TOTAL Scalar Value 5,058.00

Form Collection

#Forms 22
#Form Rows 7,829
Data Elmnts 403,044
Scalar Element Value 1.00
TOTAL Flagged Points 2,962
TOTAL Scalar Value 2,962.00

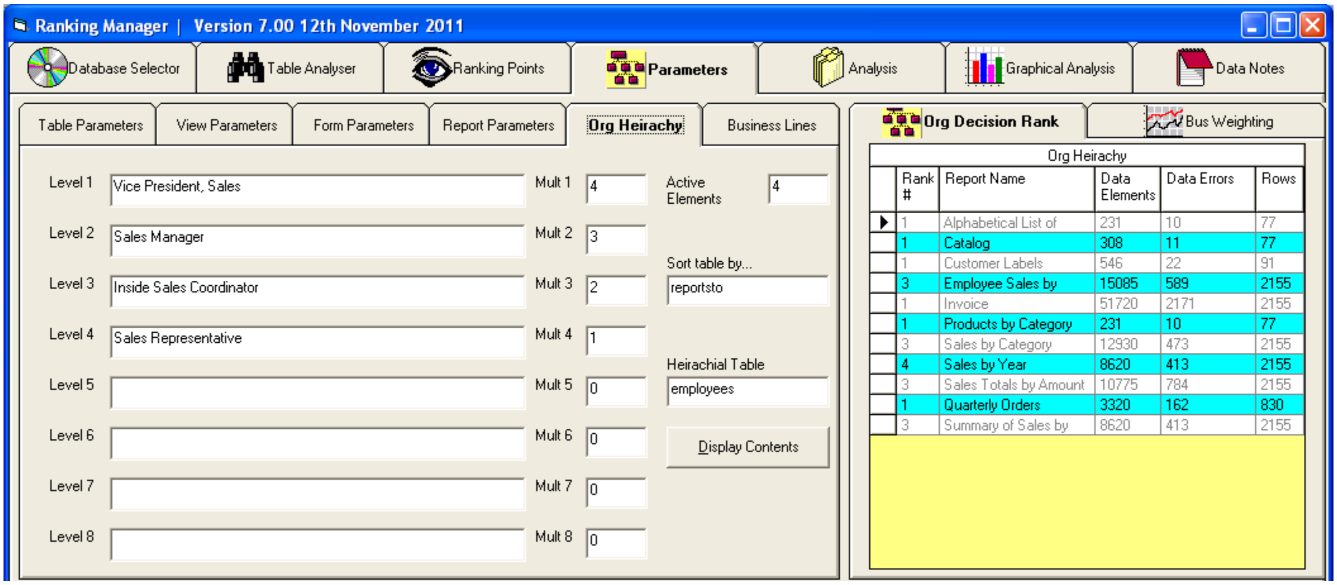
Org Heirachy

	Rank #	Report Name	Data Elements	Data Errors	Rows
▶	1	Alphabetical List of	231	10	77
	1	Catalog	308	11	77
	1	Customer Labels	546	22	91
	3	Employee Sales by	15085	589	2155
	1	Invoice	51720	2171	2155
	1	Products by	231	10	77
	3	Sales by Category	12930	473	2155
	4	Sales by Year	8620	413	2155
	3	Sales Totals by	10775	784	2155
	1	Quarterly Orders	3320	162	830
	3	Summary of Sales	8620	413	2155

Figure 6-13 5% Data Quality Error Propagation Rate (Flagged Points)

Figure 6-13 5% Data Quality Error Propagation Rate (Flagged Points) shows the baseline figures that represent the database with a 5% seeded error rate.

Report Ranking by usage where the usage numbers are represented by the hierarchical order of the report users:



Level 1	Vice President, Sales	Mult 1	4
Level 2	Sales Manager	Mult 2	3
Level 3	Inside Sales Coordinator	Mult 3	2
Level 4	Sales Representative	Mult 4	1
Level 5		Mult 5	0
Level 6		Mult 6	0
Level 7		Mult 7	0
Level 8		Mult 8	0

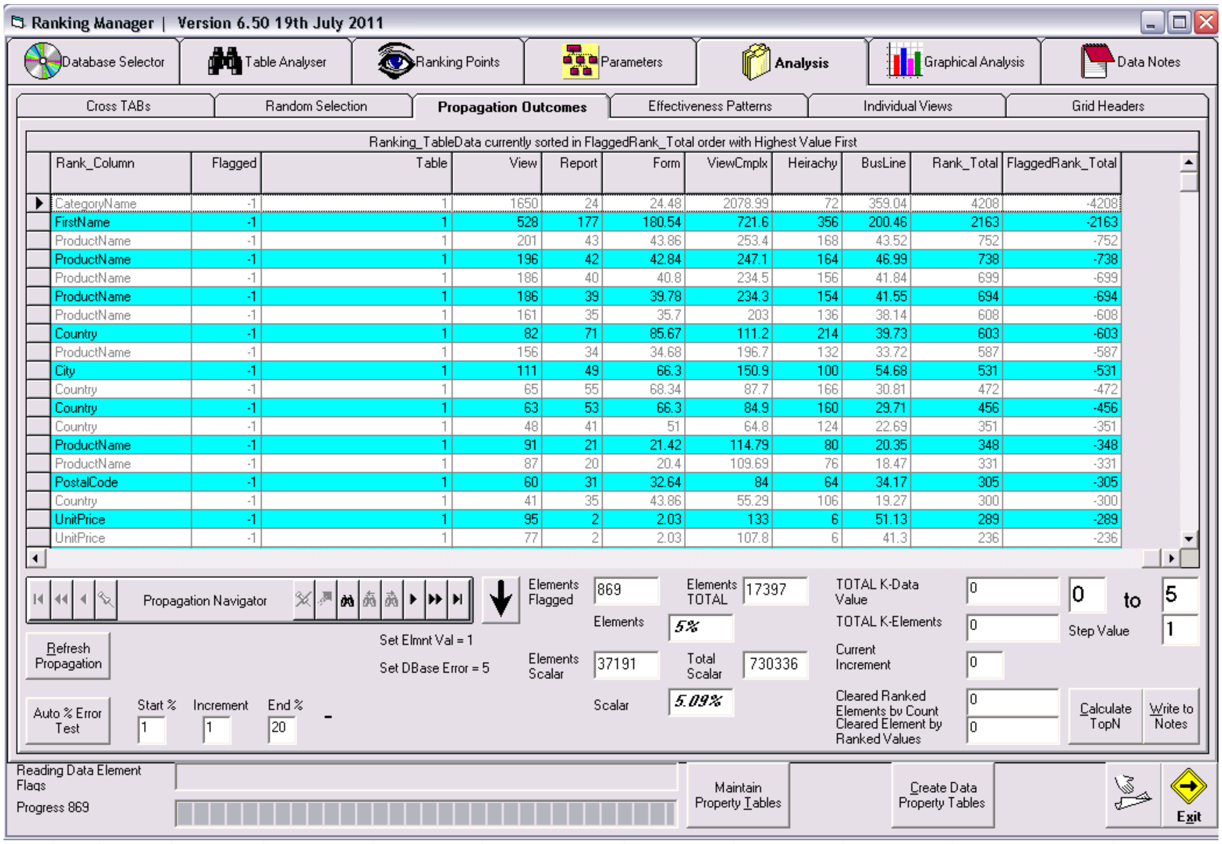
Active Elements 4
Sort table by... reportsto
Heirachial Table employees

Org Heirachy

	Rank #	Report Name	Data Elements	Data Errors	Rows
▶	1	Alphabetical List of	231	10	77
	1	Catalog	308	11	77
	1	Customer Labels	546	22	91
	3	Employee Sales by	15085	589	2155
	1	Invoice	51720	2171	2155
	1	Products by Category	231	10	77
	3	Sales by Category	12930	473	2155
	4	Sales by Year	8620	413	2155
	3	Sales Totals by Amount	10775	784	2155
	1	Quarterly Orders	3320	162	830
	3	Summary of Sales by	8620	413	2155

Figure 6-14 5% Data Quality Error by Report Hierarchical Usage

Figure 6-14 5% Data Quality Error by Report Hierarchical Usage shows the calculated error rates present in reports.



Ranking_TableData currently sorted in FlaggedRank_Total order with Highest Value First

Rank_Column	Flagged	Table	View	Report	Form	ViewCmplx	Heirachy	BusLine	Rank_Total	FlaggedRank_Total
CategoryName	-1	1	1650	24	24.48	2078.99	72	359.04	4208	-4208
FirstName	-1	1	528	177	180.54	721.6	356	200.46	2163	-2163
ProductName	-1	1	201	43	43.86	253.4	168	43.52	752	-752
ProductName	-1	1	196	42	42.84	247.1	164	46.99	738	-738
ProductName	-1	1	186	40	40.8	234.5	156	41.84	699	-699
ProductName	-1	1	186	39	39.78	234.3	154	41.55	694	-694
ProductName	-1	1	161	35	35.7	203	136	38.14	608	-608
Country	-1	1	82	71	85.67	111.2	214	39.73	603	-603
ProductName	-1	1	156	34	34.68	196.7	132	33.72	587	-587
City	-1	1	111	49	66.3	150.9	100	54.68	531	-531
Country	-1	1	65	55	68.34	87.7	166	30.81	472	-472
Country	-1	1	63	53	66.3	84.9	160	29.71	456	-456
Country	-1	1	48	41	51	64.8	124	22.69	351	-351
ProductName	-1	1	91	21	21.42	114.79	80	20.35	348	-348
ProductName	-1	1	87	20	20.4	109.69	76	18.47	331	-331
PostalCode	-1	1	60	31	32.64	84	64	34.17	305	-305
Country	-1	1	41	35	43.86	55.29	106	19.27	300	-300
UnitPrice	-1	1	95	2	2.03	133	6	51.13	289	-289
UnitPrice	-1	1	77	2	2.03	107.8	6	41.3	236	-236

Elements Flagged 869
Elements TOTAL 17397
Elements 5%
Set Elmnt Val = 1
Set DBase Error = 5
Elements Scalar 37191
Total Scalar 730336
Scalar 5.09%
TOTAL K-Data Value 0
TOTAL K-Elements 0
Current Increment 0
Cleared Ranked Elements by Count 0
Cleared Element by Ranked Values 0
0 to 5
Step Value 1
Start % 1
Increment 1
End % 20
Reading Data Element Flags
Progress 869

Figure 6-15 Control Display at 5% error rate

Table 6-7 5% error rate. Clearance rate in Unranked and Ranked order has been sorted into data elements that have data quality errors and are ranked the highest. This is the control display. This table shows that 869 data elements have been flagged as having data quality errors.

It also shows the scalar ratio for these data elements as 6081 from 730336 representing 5.09% of the total scalar value.

Table 6-7 5% error rate. Clearance rate in Unranked *and* Ranked order

Examination Effort		Unranked Data		Ranked Data		Effectiveness
Data Elements Examined %	Data Elements Examined	Cumulative Scalar Value	Cumulative Unit Count	Cumulative Scalar Value	Cumulative Unit Count	Ranked Compared *to* Unranked
1%	173	4592	5	12511	12	63
2%	347	5583	12	14435	18	61
3%	521	6138	21	15228	22	60
4%	695	7979	29	15976	28	50
5%	869	6589	35	17109	39	61

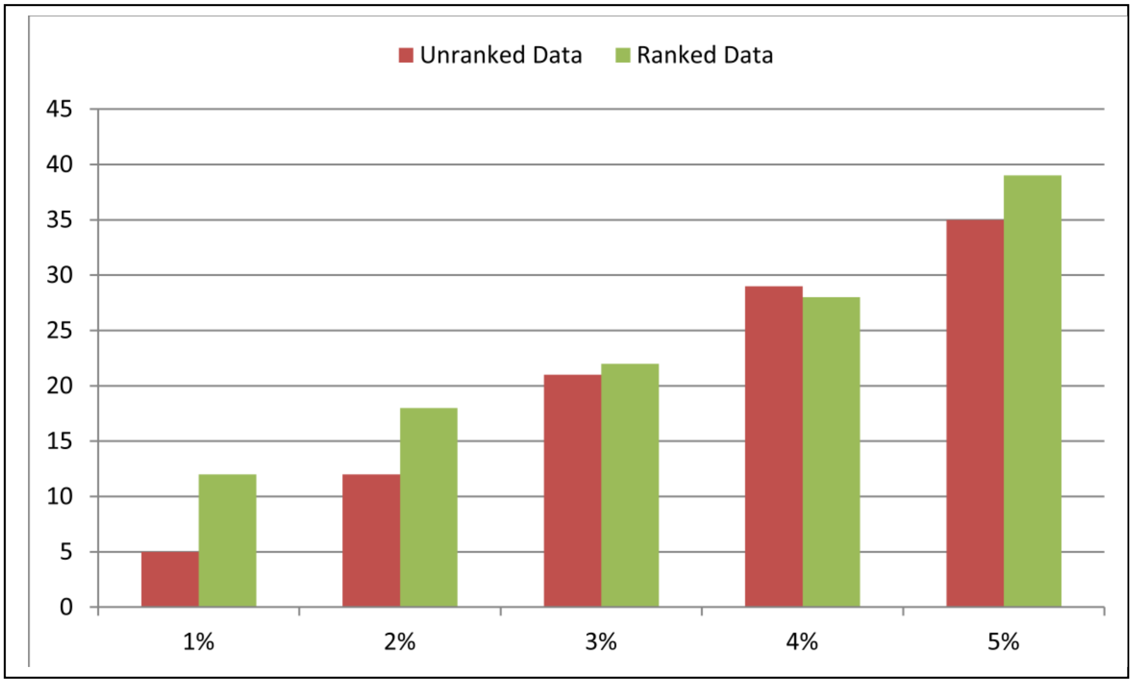


Figure 6-16 1% to 5% Unranked Vs Ranked Data - Data Units Detected

Figure 6-16 1% to 5% shows that the difference between ranked and unranked data clearance rates is negligible and, in practice, may not offer any advantages at all. This outcome is expected with the illustrated difference as function of random seeding.

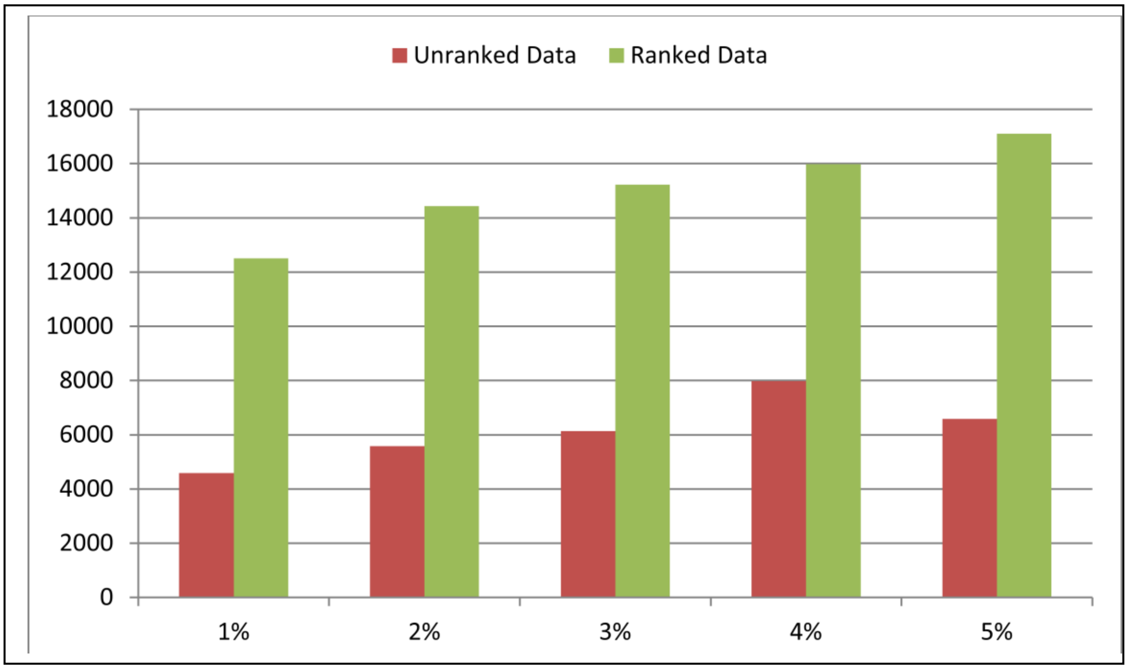


Figure 6-17 1% to 5% Unranked Vs Ranked Data - Scalar Values Detected

However, *Figure 6-17 1% to 5% Unranked Vs Ranked Data -* shows the differences between clearing unranked and ranked data elements when the success measure is the *ranked value* of the data cleared.

It is significant that the process of sorting data by ranked value also yields substantial benefits in terms of scalar rankings cleared in the first 1% to 2% of the data elements tested and cleared.

In this example at 5% introduced errors,

- clearing 1% (12/173 units) of the **ranked** data elements **clears** 12,511 / 37,191 units of scalar value; and
- clearing 1% (35/173 units) of the **unranked** data elements **clears** 4,592 / 37,191 units of scalar value.
- clearing 5% (869/869 units) of the **unranked** data elements **clears** 6,589 / 37,191 units of scalar value.

If the costs associated with data element inspection are the same regardless of the outcome, inspecting ranked data elements presents approximately twice the yield as for unranked data for around one quarter the examination effort.

The random detection approach yields very low data quality error 'hit rates' and a low scalar ranked data total.

6.4.1 *KNOWN QUALITY ERRORS POINTS USING RANKED DATA*

This is the control component where the data is ranked in scalar order and in flagged order. This shows that 5% flagged errors are cleared with 5% inspection rates.

Table 6-8 CONTROL at 5% error. Clearance rate in Ranked and Data error order

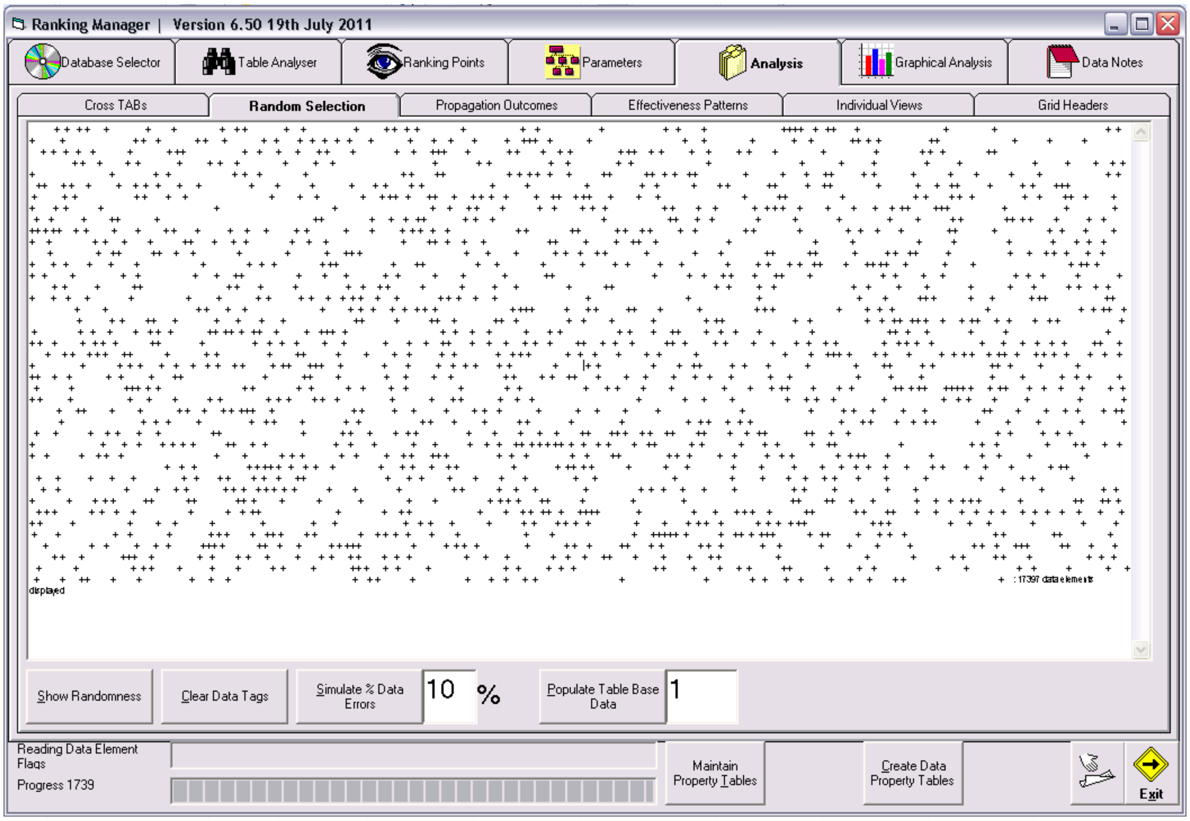
% Data Elements Examined	Data elements checked	Scalar Value	Flagged Data Units Detected
1%	173	23326	173
2%	347	28885	347
3%	521	33399	521
4%	695	36313	695
5%	869	37191	869

6.5 EXPERIMENT FINDINGS AT 5%

The scalar value of the detected data element shows a significant difference to the random scalar values although the two methods roughly match each other in terms of unit counts. This demonstrates that with 5% data quality errors, there is a significant improvement in high ranked data containing quality errors.

6.6 EXPERIMENTS USING 10% SIMULATED ERRORS

Using a 10% simulated error rate, the advantage using ranked data with unknown data elements returns a similar contestable data count, but addresses the highest ranked data elements early thus minimising the number of contestable data elements that need to be measured and then managed.



: 17397 data elements displayed

Simulate % Data Errors: 10 %

Populate Table Base Data: 1

Reading Data Element Flags

Progress 1739

Figure 6-18 10% Simulated Error Random Display

Table 6-9 10% data quality error propagation in Tables, Views, Forms and Reports

	Tables	Views	Forms	Reports
Contestable Data Elements	17397	182650	403044	719782
Flagged Elements	1739	15835	10391	5557

Report Ranking by usage where the usage numbers are represented by the hierarchical order of the report users:

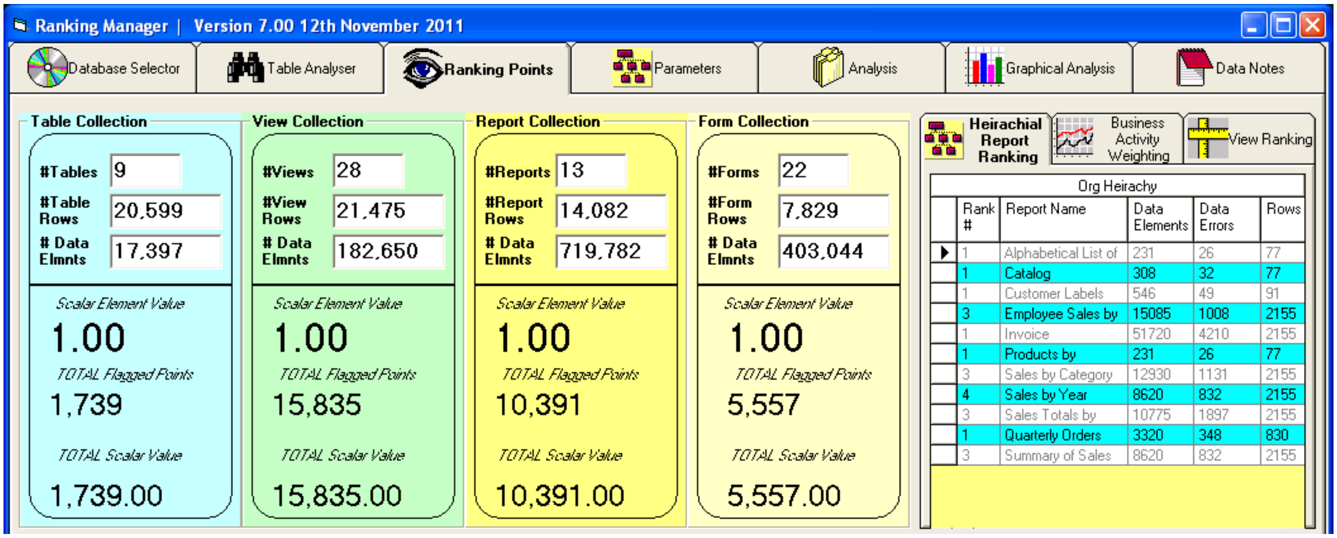


Table Collection	View Collection	Report Collection	Form Collection
#Tables 9	#Views 28	#Reports 13	#Forms 22
#Table Rows 20,599	#View Rows 21,475	#Report Rows 14,082	#Form Rows 7,829
# Data Elmnts 17,397	# Data Elmnts 182,650	# Data Elmnts 719,782	# Data Elmnts 403,044
Scalar Element Value 1.00	Scalar Element Value 1.00	Scalar Element Value 1.00	Scalar Element Value 1.00
TOTAL Flagged Points 1,739	TOTAL Flagged Points 15,835	TOTAL Flagged Points 10,391	TOTAL Flagged Points 5,557
TOTAL Scalar Value 1,739.00	TOTAL Scalar Value 15,835.00	TOTAL Scalar Value 10,391.00	TOTAL Scalar Value 5,557.00

Org Heirachy

Rank #	Report Name	Data Elements	Data Errors	Rows
1	Alphabetical List of	231	26	77
1	Catalog	308	32	77
1	Customer Labels	546	49	91
3	Employee Sales by	15085	1008	2155
1	Invoice	51720	4210	2155
1	Products by	231	26	77
3	Sales by Category	12930	1131	2155
4	Sales by Year	8620	832	2155
3	Sales Totals by	10775	1897	2155
1	Quarterly Orders	3320	348	830
3	Summary of Sales	8620	832	2155

Figure 6-19 Hierarchical Report Ranking

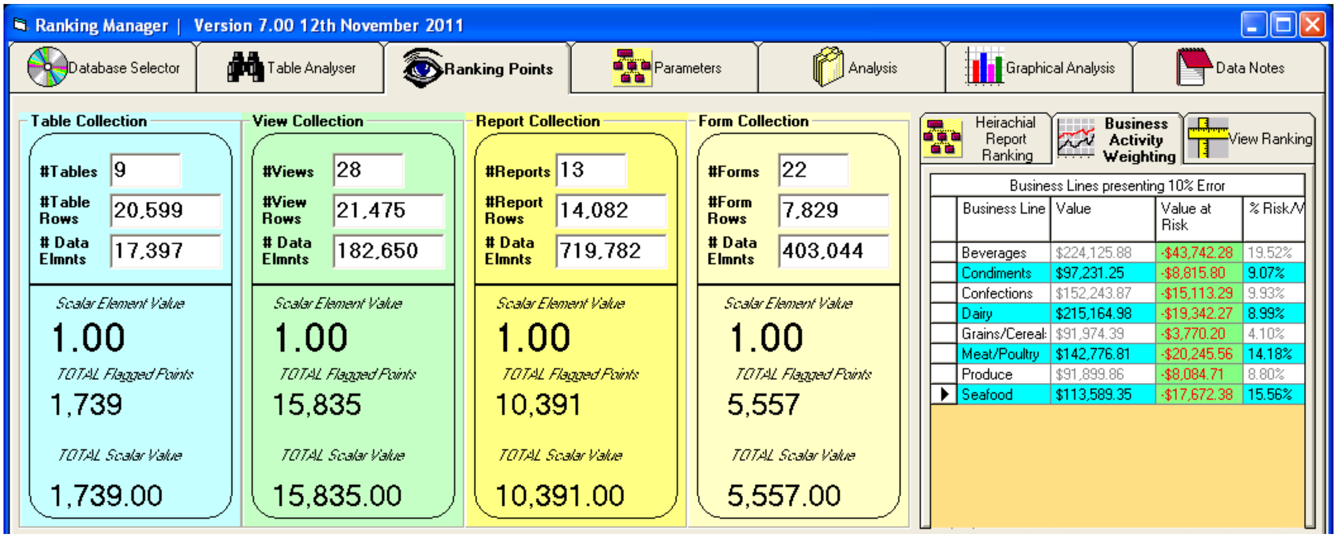


Table Collection	View Collection	Report Collection	Form Collection
#Tables 9	#Views 28	#Reports 13	#Forms 22
#Table Rows 20,599	#View Rows 21,475	#Report Rows 14,082	#Form Rows 7,829
# Data Elmnts 17,397	# Data Elmnts 182,650	# Data Elmnts 719,782	# Data Elmnts 403,044
Scalar Element Value 1.00	Scalar Element Value 1.00	Scalar Element Value 1.00	Scalar Element Value 1.00
TOTAL Flagged Points 1,739	TOTAL Flagged Points 15,835	TOTAL Flagged Points 10,391	TOTAL Flagged Points 5,557
TOTAL Scalar Value 1,739.00	TOTAL Scalar Value 15,835.00	TOTAL Scalar Value 10,391.00	TOTAL Scalar Value 5,557.00

Business Lines presenting 10% Error

Business Line	Value	Value at Risk	% Risk/V
Beverages	$224,125.88	-$43,742.28	19.52%
Condiments	$97,231.25	-$8,815.80	9.07%
Confections	$152,243.87	-$15,113.29	9.93%
Dairy	$215,164.98	-$19,342.27	8.99%
Grains/Cereal:	$91,974.39	-$3,770.20	4.10%
Meat/Poultry	$142,776.81	-$20,245.56	14.18%
Produce	$91,899.86	-$8,084.71	8.80%
Seafood	$113,589.35	-$17,672.38	15.56%

Figure 6-20 Business Lines showing 'values at risk'

6.7 Experiment Findings at 10%

Note that at 10% the actual 'at risk' percentages are tending to better match the seeded error rate more closely than when seeded at 1%.

The exception in this case is the 'Grains/Cereals' which presents lower 'at risk' amount. This too illustrates the variations that data quality errors can exhibit with different views, reporting, usage and value against product lines.

Table 6-10 Value at risk for 10% error simulation

Sales Category	Total Value	Value at Risk	% Value at Risk
Beverages	$243,933.24	-$23,934.92	09.81%
Condiments	$ 94,662.58	-$11,384.47	12.03%
Confections	$148,262.63	-$19,094.53	12.88%
Dairy Products	$225,098.92	-$ 9,408.33	04.18%
Grains/Cereals	$ 94,158.59	-$ 1,586.00	01.68%
Meat/Poultry	$147,124.18	-$15,898.19	10.81%
Produce	$ 90,848.56	-$ 9,136.01	10.06%
Seafood	$121,791.99	-$ 9,469.74	07.78%

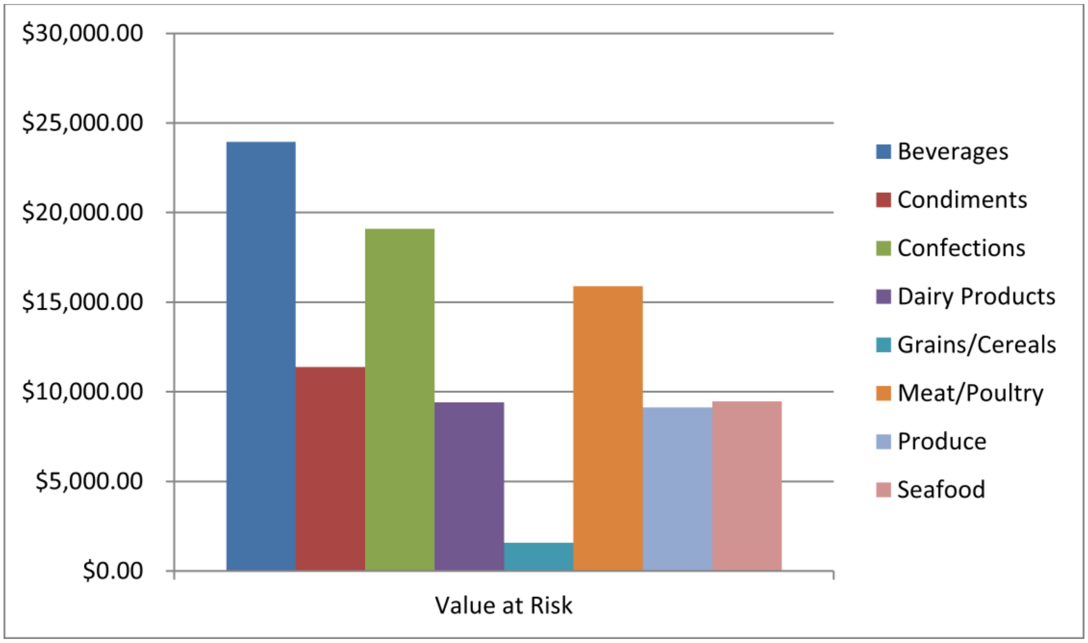


Figure 6-21 10% error simulation for Total Product Value

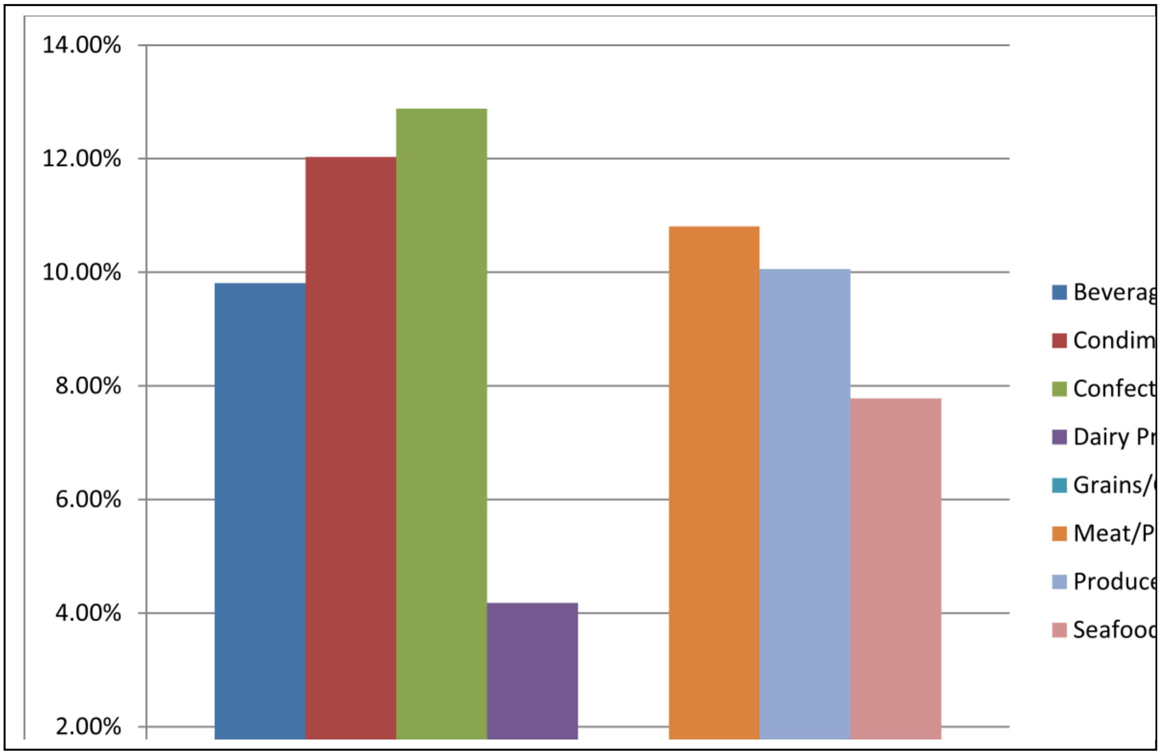


Figure 6-22 10% error simulation for product % Value at Risk

6.7.1 *Product Line Summation*

The variation between product lines showing ‘Value at Risk’ illustrates the notion that data quality errors can effect categories in a roughly predicable manner; with the potential for variations based on product value quantity, unit price, units traded and the potential for some data quality errors to be prolific.

The value to decision makers is when considering decisions or evaluating performance indicators.

Notable is that different business lines do not reflect the 10% error rate due to different volumes and price per unit variations.

The ‘Value at Risk’ is a representation of the derived values where at least one of the data elements in each formula contains a flagged data element.

The graph below shows the three methods at 10% simulated errors at each percentage clearance rate from 1% through to 10%.

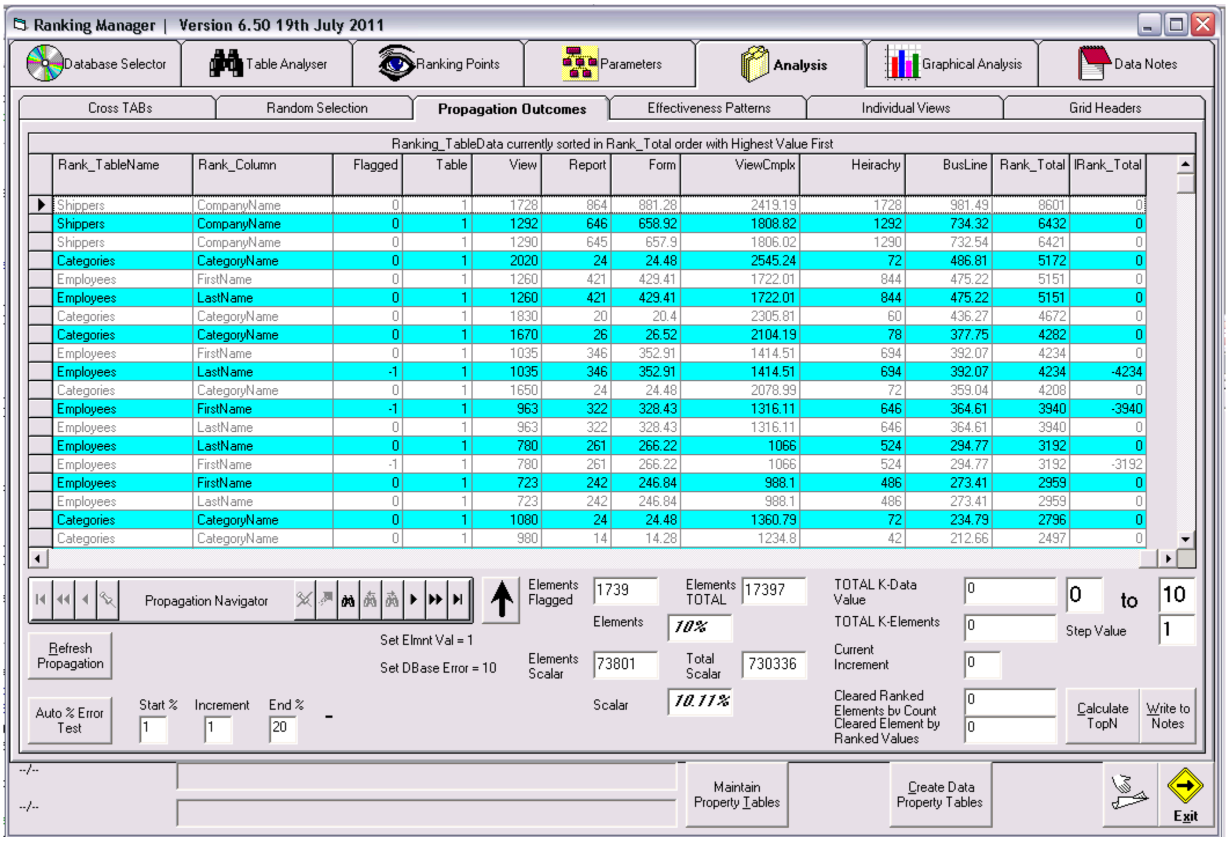
The simulation for 10% data quality errors yields 1739 flagged data elements from 17397 in total.

Log Book entry:

Database: Northwind.mdb. Contains 17397 contestable data elements. These elements have been tagged representing a 10% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8914

The ranked table and data element flags are shown in *Figure 6-23 Ranked Table and Data Element Flags.*

This table shows that 1739 data elements from 17397 have been flagged as having data quality errors. The process is to ‘clear’ various proportions of the data errors in increments of 1% from 1% through to 10% to see the difference between random detection, ranked detection and (the control) ranked detection with data errors known.



Ranking_TableData currently sorted in Rank_Total order with Highest Value First

Rank_TableName	Rank_Column	Flagged	Table	View	Report	Form	ViewCmplx	Heirachy	BusLine	Rank_Total	IRank_Total
Shippers	CompanyName	0	1	1728	864	881.28	2419.19	1728	981.49	8601	0
Shippers	CompanyName	0	1	1292	646	658.92	1808.82	1292	734.32	6432	0
Shippers	CompanyName	0	1	1290	645	657.9	1806.02	1290	732.54	6421	0
Categories	CategoryName	0	1	2020	24	24.48	2545.24	72	486.81	5172	0
Employees	FirstName	0	1	1260	421	429.41	1722.01	844	475.22	5151	0
Employees	LastName	0	1	1260	421	429.41	1722.01	844	475.22	5151	0
Categories	CategoryName	0	1	1830	20	20.4	2305.81	60	436.27	4672	0
Categories	CategoryName	0	1	1670	26	26.52	2104.19	78	377.75	4282	0
Employees	FirstName	0	1	1035	346	352.91	1414.51	694	392.07	4234	0
Employees	LastName	-1	1	1035	346	352.91	1414.51	694	392.07	4234	-4234
Categories	CategoryName	0	1	1650	24	24.48	2078.99	72	359.04	4208	0
Employees	FirstName	-1	1	963	322	328.43	1316.11	646	364.61	3940	-3940
Employees	LastName	0	1	963	322	328.43	1316.11	646	364.61	3940	0
Employees	LastName	0	1	780	261	266.22	1066	524	294.77	3192	0
Employees	FirstName	-1	1	780	261	266.22	1066	524	294.77	3192	-3192
Employees	FirstName	0	1	723	242	246.84	988.1	486	273.41	2959	0
Employees	LastName	0	1	723	242	246.84	988.1	486	273.41	2959	0
Categories	CategoryName	0	1	1080	24	24.48	1360.79	72	234.79	2796	0
Categories	CategoryName	0	1	980	14	14.28	1234.8	42	212.66	2497	0

Elements Flagged 1739
Elements TOTAL 17397
Elements 10%
Set Elmnt Val = 1
Set DBase Error = 10
Elements Scalar 73801
Total Scalar 730336
Scalar 10.11%
TOTAL K-Data Value 0
TOTAL K-Elements 0
Current Increment 0
Cleared Ranked Elements by Count 0
Cleared Element by Ranked Values 0
0 to 10
Step Value 1
Start % 1
Increment 1
End % 20

Figure 6-23 Ranked Table and Data Element Flags

Table 6-11 10% error rate. Clearance rate in Unranked *and* Ranked order

***Remediation* effort**		**Unranked Data**		**Ranked Data**		**Effectiveness**
Data Elements Examined %	**Data Elements Examined**	**Cumulative Scalar Value**	**Cumulative Unit Count**	**Cumulative Scalar Value**	**Cumulative Unit Count**	**Ranked Compared *to* Unranked**
1%	173	**2888**	19	**20914**	16	86%
2%	347	**6194**	32	**28363**	38	78%
3%	521	**6820**	41	**31350**	52	78%
4%	695	**8088**	56	**33350**	67	76%
5%	869	**9191**	71	**35415**	87	74%
6%	1043	**9735**	88	**36828**	105	74%
7%	1217	**10343**	107	**38273**	124	73%
8%	1391	**11015**	128	**39729**	147	72%
9%	1565	**11399**	140	**40698**	166	72%
10%	1739	**11783**	152	**41973**	191	72%

Figure 6-24 1% to 10% Unranked Vs Ranked Data - Data Elements Cleared shows that the difference between ranked and unranked data clearance rates is negligible and, in practice, may not offer any advantages at all. The illustrated difference is a function of random seeding.

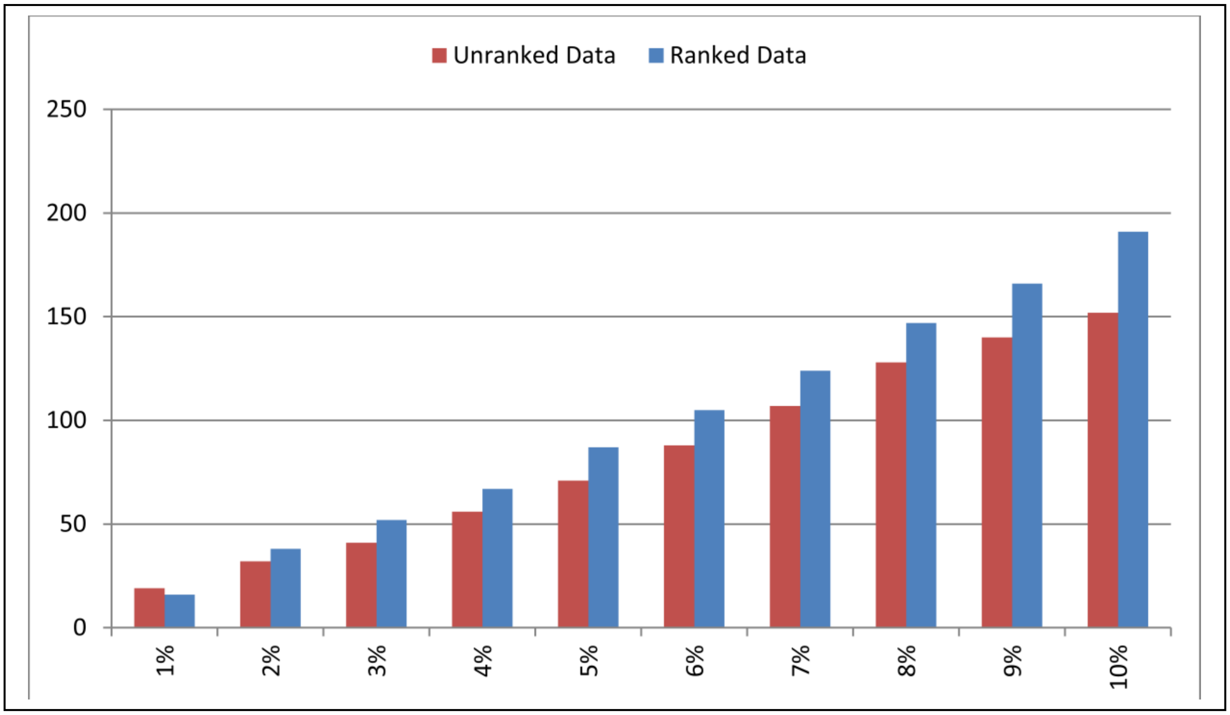


Figure 6-24 1% to 10% Unranked Vs Ranked Data - Data Elements Cleared

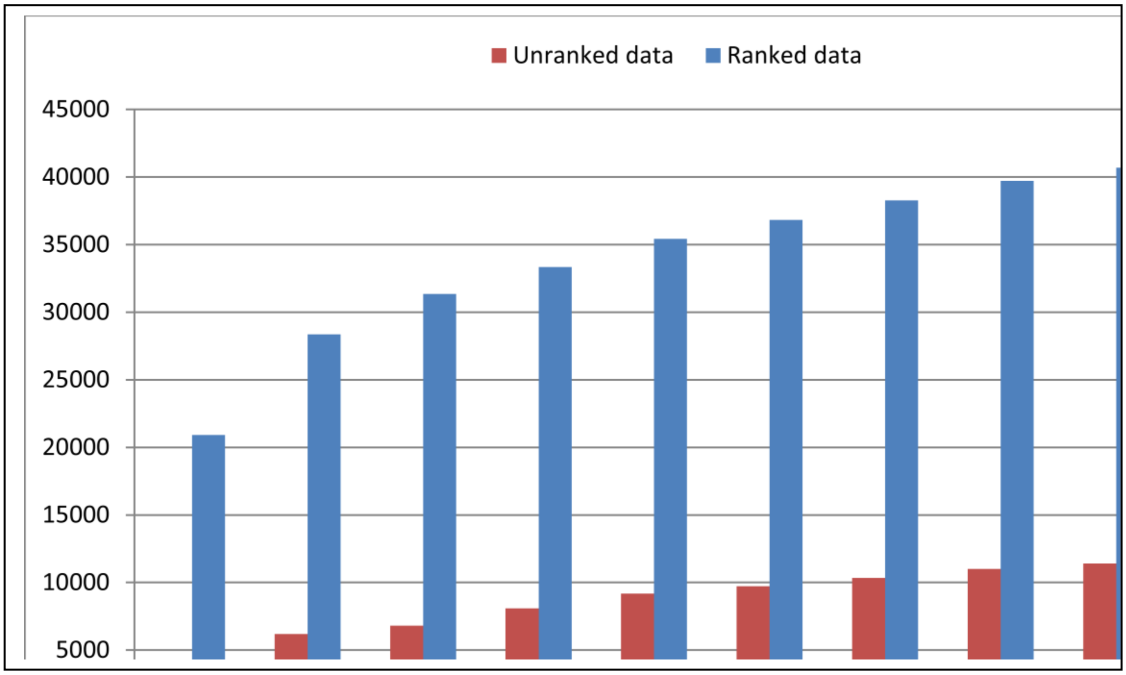


Figure 6-25 1% to 10% Unranked Vs Ranked Data - Ranked Values Cleared

Figure 6-25 1% to 10% Unranked Vs Ranked Data - Ranked Values Cleared shows the differences between clearing unranked and ranked data elements when the success measure is the ranked value of the data cleared.

Of greater significance is that the process of sorting data by ranked value also yields substantial benefits in terms of scalar rankings cleared in the first 1% of the data elements tested and cleared.

In this example at 10% introduced errors,

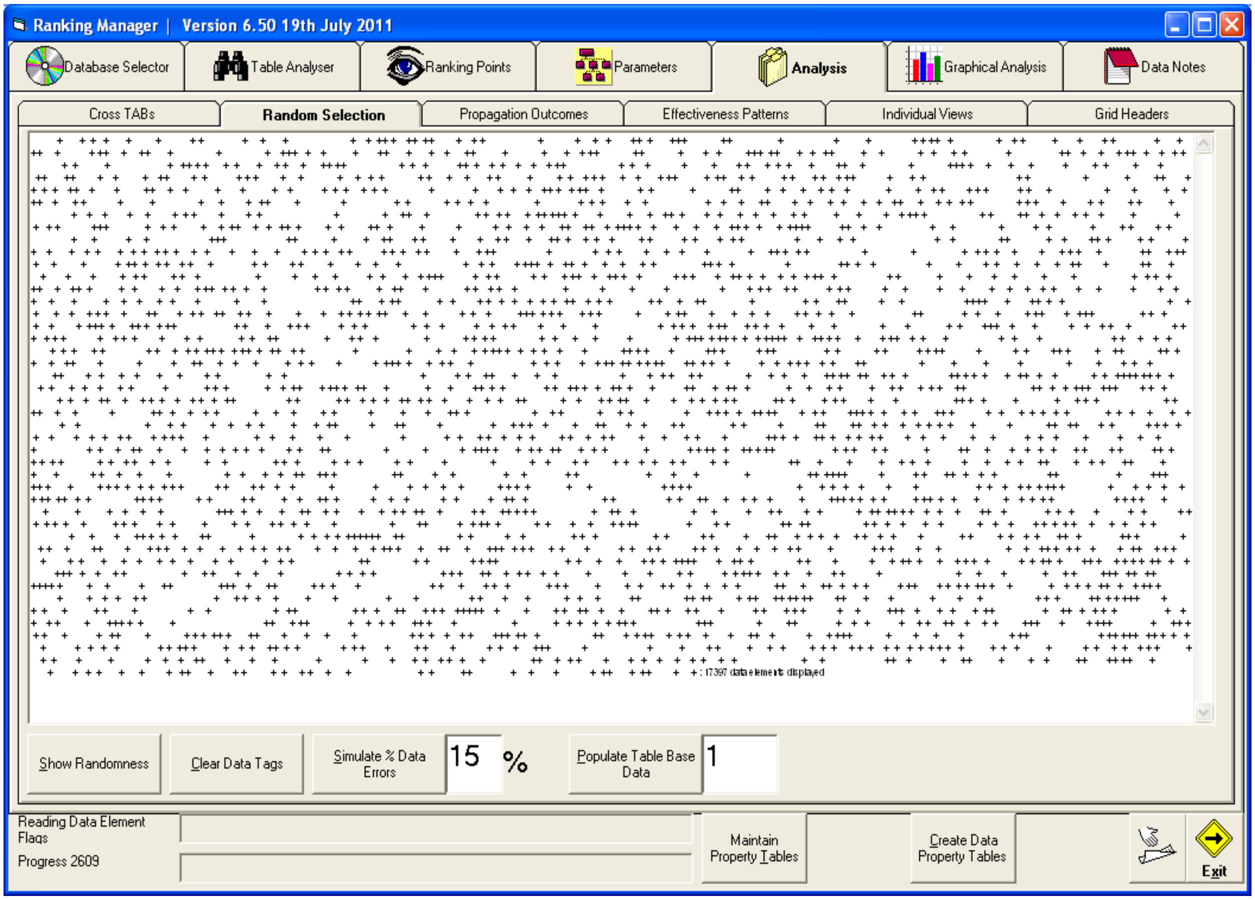
- Clearing 1% (173 units) of the **ranked** data elements clears 20,914 units of scalar value
- Clearing 1% (173 units) of the **unranked** data elements clears 2,888 units of scalar value.
- Clearing 10% (1739 units) of the **unranked** data elements clears 11,783 units of scalar value.

Assuming that costs associated with data element inspection are the same regardless of the outcome, inspecting ranked data elements presents slightly over twice the yield as for unranked data for one tenth the examination effort.

Clearing 173 data elements in ranked order offers a significant seven times advantage over unranked data element inspection.

6.8 Experiments Using 15% Simulated Errors

Using a 15% simulated error rate, the advantage using ranked data with unknown data elements returns a similar contestable data count, but addresses the highest ranked data elements early thus minimising the number of contestable data elements that need to be measured and then managed.



: 17397 data elements displayed

Simulate % Data Errors: 15 %

Populate Table Base Data: 1

Reading Data Element Flags

Progress 2609

Figure 6-26 15% Simulated Error Random Display

Table 6-12 15% data quality error propagation in Tables, Views, Forms and Reports

	Tables	Views	Forms	Reports
Contestable Data Elements	17397	182650	403044	719782
Flagged Elements	2609	23090	15037	8512

Report Ranking by usage where the usage numbers are represented by the hierarchical order of the report users:

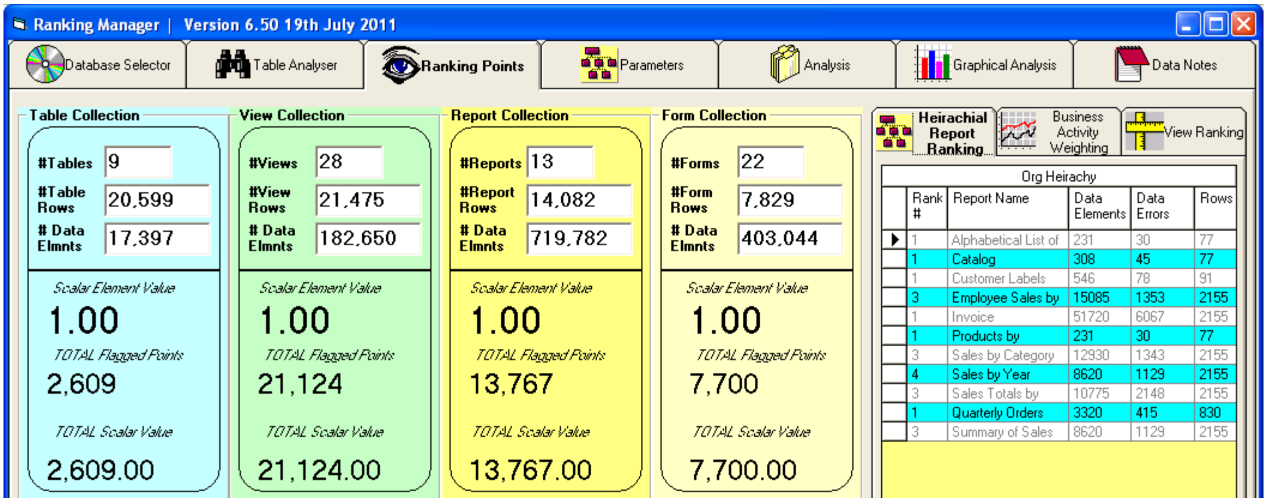


	Table Collection	View Collection	Report Collection	Form Collection
	#Tables 9	#Views 28	#Reports 13	#Forms 22
	#Table Rows 20,599	#View Rows 21,475	#Report Rows 14,082	#Form Rows 7,829
	# Data Elmnts 17,397	# Data Elmnts 182,650	# Data Elmnts 719,782	# Data Elmnts 403,044
Scalar Element Value	1.00	1.00	1.00	1.00
TOTAL Flagged Points	2,609	21,124	13,767	7,700
TOTAL Scalar Value	2,609.00	21,124.00	13,767.00	7,700.00

Org Heirachy

Rank #	Report Name	Data Elements	Data Errors	Rows
1	Alphabetical List of	231	30	77
1	Catalog	308	45	77
1	Customer Labels	546	78	91
3	Employee Sales by	15085	1353	2155
1	Invoice	51720	6067	2155
1	Products by	231	30	77
3	Sales by Category	12930	1343	2155
4	Sales by Year	8620	1129	2155
3	Sales Totals by	10775	2148	2155
1	Quarterly Orders	3320	415	830
3	Summary of Sales	8620	1129	2155

Figure 6-27 Hierarchical Report Ranking

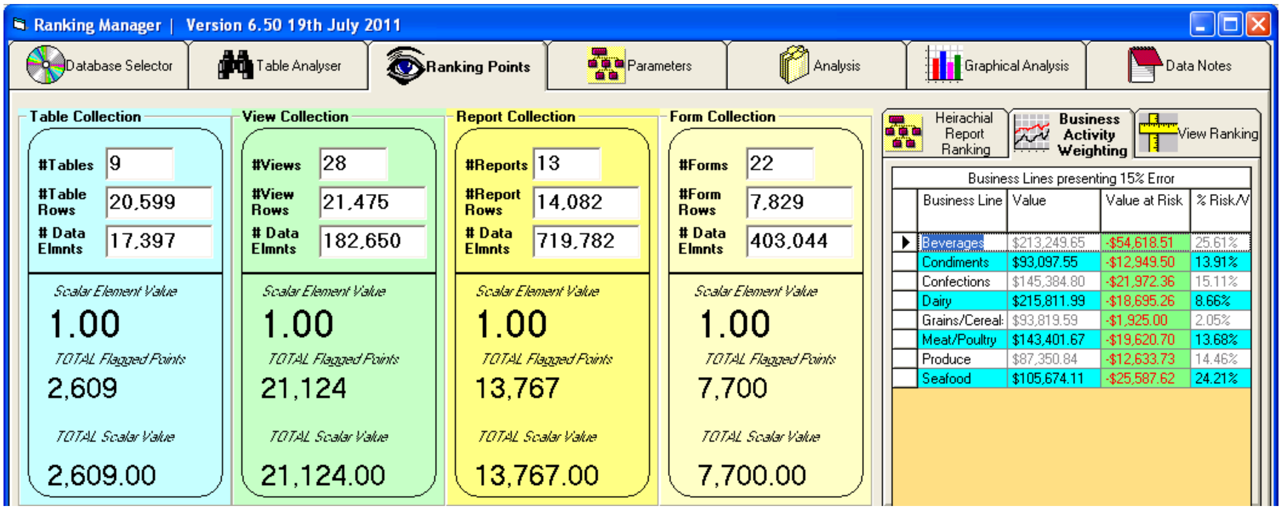


	Table Collection	View Collection	Report Collection	Form Collection
	#Tables 9	#Views 28	#Reports 13	#Forms 22
	#Table Rows 20,599	#View Rows 21,475	#Report Rows 14,082	#Form Rows 7,829
	# Data Elmnts 17,397	# Data Elmnts 182,650	# Data Elmnts 719,782	# Data Elmnts 403,044
Scalar Element Value	1.00	1.00	1.00	1.00
TOTAL Flagged Points	2,609	21,124	13,767	7,700
TOTAL Scalar Value	2,609.00	21,124.00	13,767.00	7,700.00

Business Lines presenting 15% Error

Business Line	Value	Value at Risk	% Risk/V
Beverages	$213,249.65	-$54,618.51	25.61%
Condiments	$93,097.55	-$12,949.50	13.91%
Confections	$145,384.80	-$21,972.36	15.11%
Dairy	$215,811.99	-$18,695.26	8.66%
Grains/Cereal:	$93,819.59	-$1,925.00	2.05%
Meat/Poultry	$143,401.67	-$19,620.70	13.68%
Produce	$87,350.84	-$12,633.73	14.46%
Seafood	$105,674.11	-$25,587.62	24.21%

Figure 6-28 Business Lines showing 'values at risk'

6.9 Experiment Findings at 15%

Note that at 15% the actual ‘at risk’ percentages are tending to better match the seeded error rate more closely than when seeded at 1%.

The exception in this case is the ‘Grains/Cereals’ which presents lower ‘at risk’ amount. This too illustrates the variations that data quality errors can exhibit with different views, reporting, usage and value against product lines.

Table 6-13 Value at risk for 15% error simulation

Sales Category	Total Value	Value at Risk	% Value at Risk
Beverages	$213,249.65	$54,618.51	25.61%
Condiments	$ 93,097.55	$12,949.50	13.91%
Confections	$145,384.80	$21,972.36	15.11%
Dairy Products	$215,811.99	$18,695.26	8.66%
Grains/Cereals	$ 93,819.59	$ 1,925.00	2.05%
Meat/Poultry	$143,401.67	$19,620.70	13.68%
Produce	$ 87,350.84	$12,633.73	14.46%
Seafood	$105,674.11	$25,587.62	24.21%

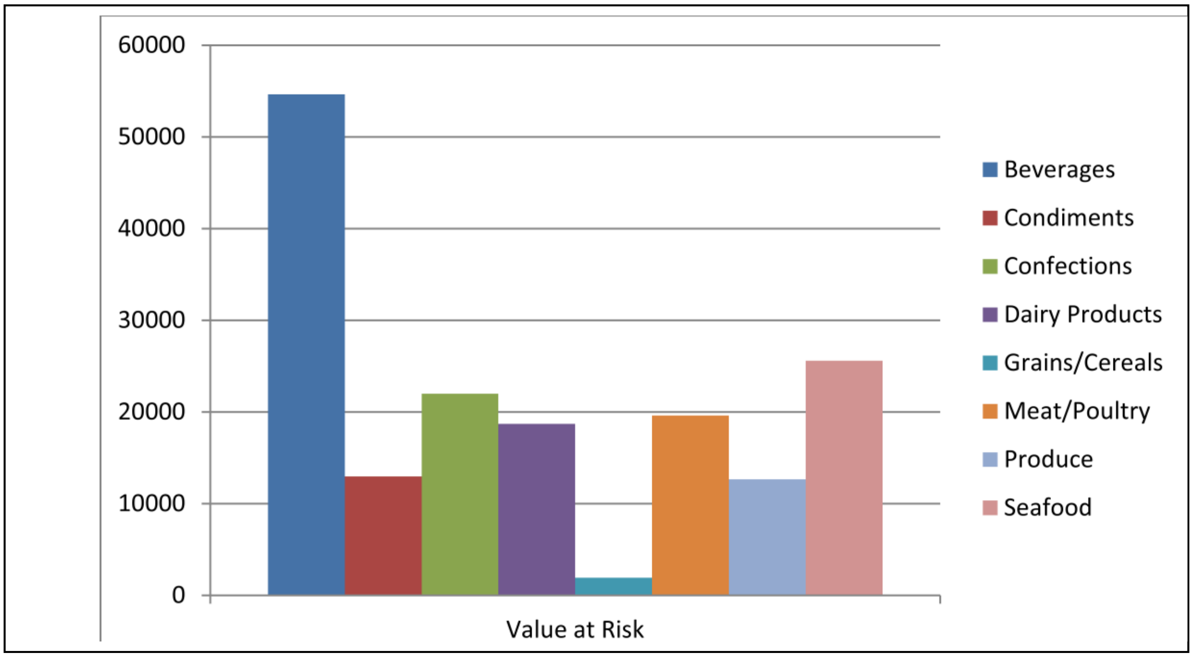


Figure 6-29 15% error simulation for Total Product Value

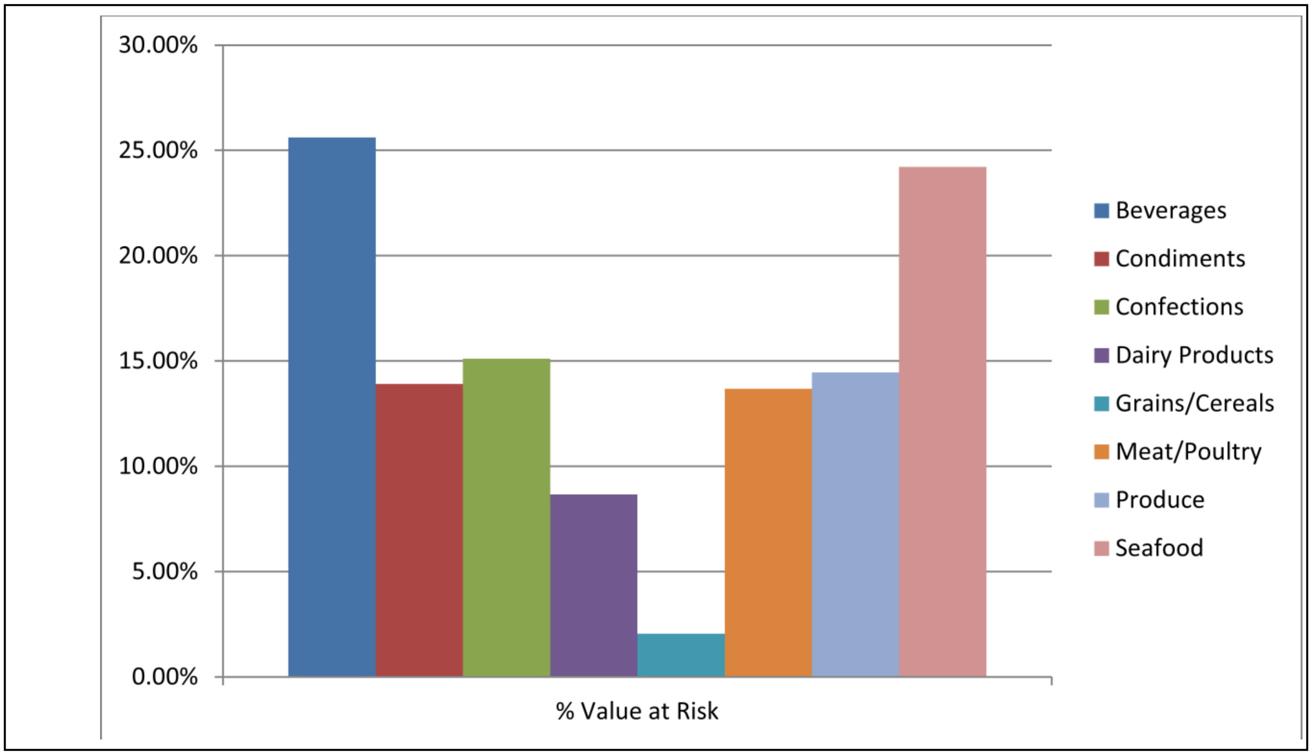


Figure 6-30 15% error simulation for product % Value at Risk

6.9.1 PRODUCT LINE SUMMATION

The variation between product lines showing ‘Value at Risk’ illustrates the notion that data quality errors can affect categories in a roughly predicable manner; with the potential for variations based on product value quantity, unit price, units traded and the potential for some data quality errors to be prolific.

The value to decision makers is when considering decisions or evaluating performance indicators.

Notable is that different business lines do not reflect the 10% error rate due to different volumes and price per unit variations.

The ‘Value at Risk’ is a representation of the derived values where at least one of the data elements in each formula contains a flagged data element.

The graph below shows the three methods at 15% simulated errors at each percentage clearance rate from 1% through to 15%.

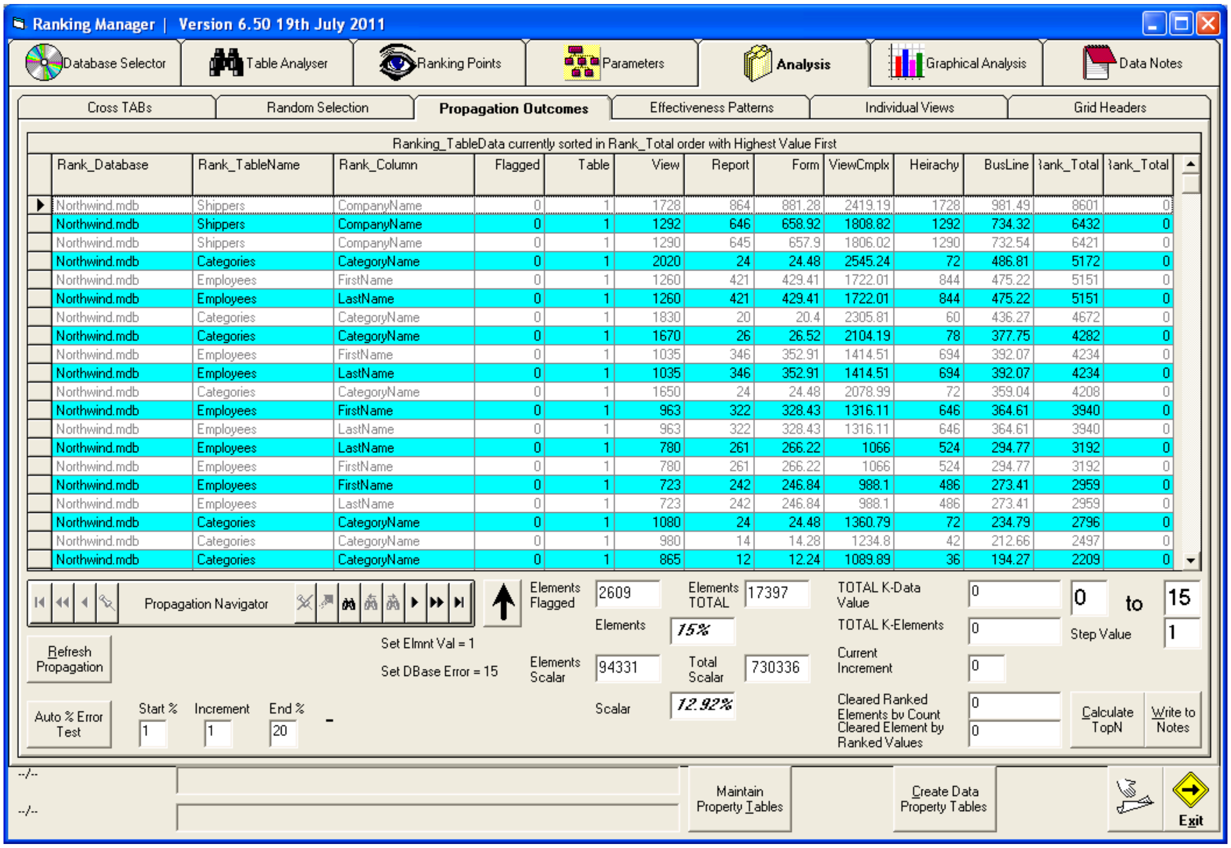
The simulation for 15% data quality errors yields 1739 flagged data elements from 17397 in total.

Log Book entry:

Database: Northwind.mdb. Contains 17397 contestable data elements. These elements have been tagged representing a 15% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8914

The ranked table and data element flags are shown in *Figure 6-23 Ranked Table and Data Element Flags.*

This table shows that 2609 data elements from 17397 have been flagged as having data quality errors. The process is to ‘clear’ various proportions of the data errors in increments of 1% from 1% through to 15% to see the difference between random detection, ranked detection and (the control) ranked detection with data errors known.



Ranking_TableData currently sorted in Rank_Total order with Highest Value First

Rank_Database	Rank_TableName	Rank_Column	Flagged	Table	View	Report	Form	ViewCmplx	Heirachy	BusLine	ỉank_Total	ỉank_Total
Northwind.mdb	Shippers	CompanyName	0	1	1728	864	881.28	2419.19	1728	981.49	8601	0
Northwind.mdb	Shippers	CompanyName	0	1	1292	646	658.92	1808.82	1292	734.32	6432	0
Northwind.mdb	Shippers	CompanyName	0	1	1290	645	657.9	1806.02	1290	732.54	6421	0
Northwind.mdb	Categories	CategoryName	0	1	2020	24	24.48	2545.24	72	486.81	5172	0
Northwind.mdb	Employees	FirstName	0	1	1260	421	429.41	1722.01	844	475.22	5151	0
Northwind.mdb	Employees	LastName	0	1	1260	421	429.41	1722.01	844	475.22	5151	0
Northwind.mdb	Categories	CategoryName	0	1	1830	20	20.4	2305.81	60	436.27	4672	0
Northwind.mdb	Categories	CategoryName	0	1	1670	26	26.52	2104.19	78	377.75	4282	0
Northwind.mdb	Employees	FirstName	0	1	1035	346	352.91	1414.51	694	392.07	4234	0
Northwind.mdb	Employees	LastName	0	1	1035	346	352.91	1414.51	694	392.07	4234	0
Northwind.mdb	Categories	CategoryName	0	1	1650	24	24.48	2078.99	72	359.04	4208	0
Northwind.mdb	Employees	FirstName	0	1	963	322	328.43	1316.11	646	364.61	3940	0
Northwind.mdb	Employees	LastName	0	1	963	322	328.43	1316.11	646	364.61	3940	0
Northwind.mdb	Employees	LastName	0	1	780	261	266.22	1066	524	294.77	3192	0
Northwind.mdb	Employees	FirstName	0	1	780	261	266.22	1066	524	294.77	3192	0
Northwind.mdb	Employees	FirstName	0	1	723	242	246.84	988.1	486	273.41	2959	0
Northwind.mdb	Employees	LastName	0	1	723	242	246.84	988.1	486	273.41	2959	0
Northwind.mdb	Categories	CategoryName	0	1	1080	24	24.48	1360.79	72	234.79	2796	0
Northwind.mdb	Categories	CategoryName	0	1	980	14	14.28	1234.8	42	212.66	2497	0
Northwind.mdb	Categories	CategoryName	0	1	865	12	12.24	1089.89	36	194.27	2209	0

Elements Flagged 2609
Elements TOTAL 17397
TOTAL K-Data Value 0
0 to 15
Elements 15%
TOTAL K-Elements 0
Step Value 1
Set Elmnt Val = 1
Set DBase Error = 15
Elements Scalar 94331
Total Scalar 730336
Current Increment 0
Start % 1
Increment 1
End % 20
Scalar 12.92%
Cleared Ranked Elements by Count 0
Cleared Element by Ranked Values 0
--/--
--/--

Figure 6-31 Ranked Table and Data Element Flags

Table 6-14 15% error rate. Clearance rate in Unranked *and* Ranked order

***Remediation* effort**		**Unranked Data**		**Ranked Data**		**Effectiveness**
Data Elements Examined %	**Data Elements Examined**	**Cumulative Scalar Value**	**Cumulative Unit Count**	**Cumulative Scalar Value**	**Cumulative Unit Count**	**Ranked Compared *to* Unranked**
1	173	**3621**	20	**14086**	22	74%
2	347	**8905**	44	**24186**	52	63%
3	521	**11810**	78	**30139**	80	61%
4	695	**14616**	105	**33204**	103	56%
5	869	**15245**	123	**36190**	132	58%
6	1043	**15949**	145	**37769**	152	58%
7	1217	**16941**	176	**39668**	177	57%
8	1391	**17549**	195	**41685**	209	58%
9	1565	**17997**	209	**43321**	241	58%
10	1739	**19085**	243	**44596**	266	57%
11	1913	**19853**	267	**45920**	293	57%
12	2087	**20781**	296	**46964**	316	56%
13	2261	**21837**	329	**47861**	339	54%
14	2435	**22669**	355	**48801**	367	54%
15	2609	**23501**	381	**49473**	388	52%

Figure 6-24 1% to 10% Unranked Vs Ranked Data - Data Elements Cleared shows that the difference between ranked and unranked data clearance rates is negligible and, in practice, does not offer an advantage. The illustrated difference is a function of random seeding.

The key difference is the rate at which the scalar data values are accumulated for the same examination effort with ranked and unranked data.

The most effective percentage examination appears (in this example) to be between 3% and 5%

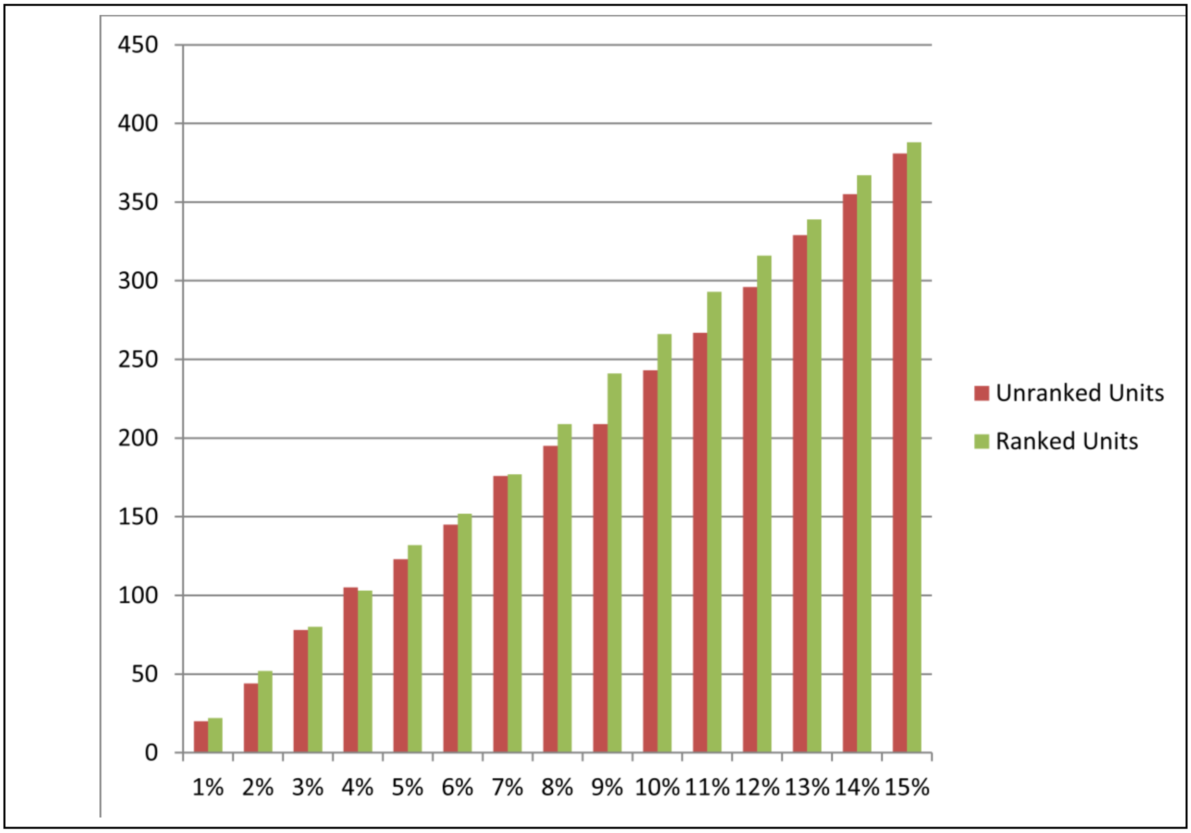


Figure 6-32 1% to 15% Unranked Vs Ranked Data - Data Elements Cleared

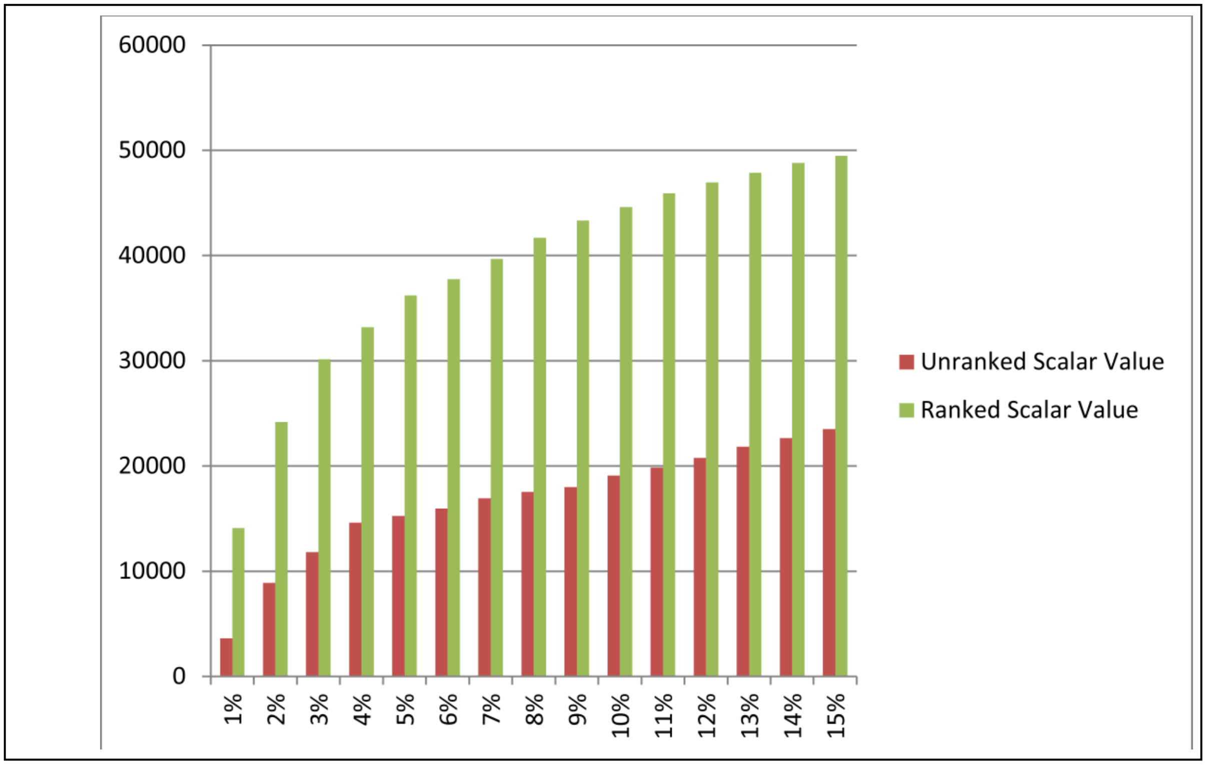


Figure 6-33 1% to 15% Unranked Vs Ranked Scalar Values shows the differences between clearing unranked and ranked data elements when the success measure is the ranked value of the data cleared.

Of greater significance is that the process of sorting data by ranked value also yields substantial benefits in terms of scalar rankings cleared in the first 1% of the data elements tested and cleared.

In this example at 10% introduced errors,

- clearing 1% (173 units) of the **ranked** data elements clears 20,914 units of scalar value
- clearing 1% (173 units) of the **unranked** data elements clears 2,888 units of scalar value.
- clearing 10% (1739 units) of the **unranked** data elements clears 11,783 units of scalar value.

Assuming that costs associated with data element inspection are the same regardless of the outcome, inspecting ranked data elements presents slightly over twice the yield as for unranked data for one tenth the examination effort.

Clearing 173 data elements in ranked order offers a seven time advantage over unranked data element inspection.

6.10 EXPERIMENT OUTCOMES

The structure, composition and content volume of a database determines the characteristic of the scalar value of the contestable data elements.

The conclusion reached in these experiments is dependent upon the structure, composition and content volume of the North Wind Traders database.

The benefit using ranked contestable data elements presents significant benefits in terms of corporate benefits for data quality management if the error rate is between 3% and 15%.

The pattern of ranked data changes quite dramatically and presents a guide for organisations to determine the target limit for data quality evaluation (and possible remediation).

The North Wind Traders database presents a curve for the top ranked 500 contestable data elements as shown below:

The North Wind Traders database shows a significant ‘drop off’ in ranked values is substantial after the first 100 ranked data elements (shown with a red oval) and may represent some return-on-investment cut-off point as the ranked value drops from over 8000 to less than 1000.

When graphed with the seeded data quality errors, the benefit of the ranked data quality evaluation becomes clear in the pivot graph *Figure 6-34 10% Simulated Data Quality errors*. This graph shows that the data element ranking in ranking order reduce to low scalar values around the 500th (from over 17000) data elements) ranked point allowing a planned remediation target to be assessed and implemented.

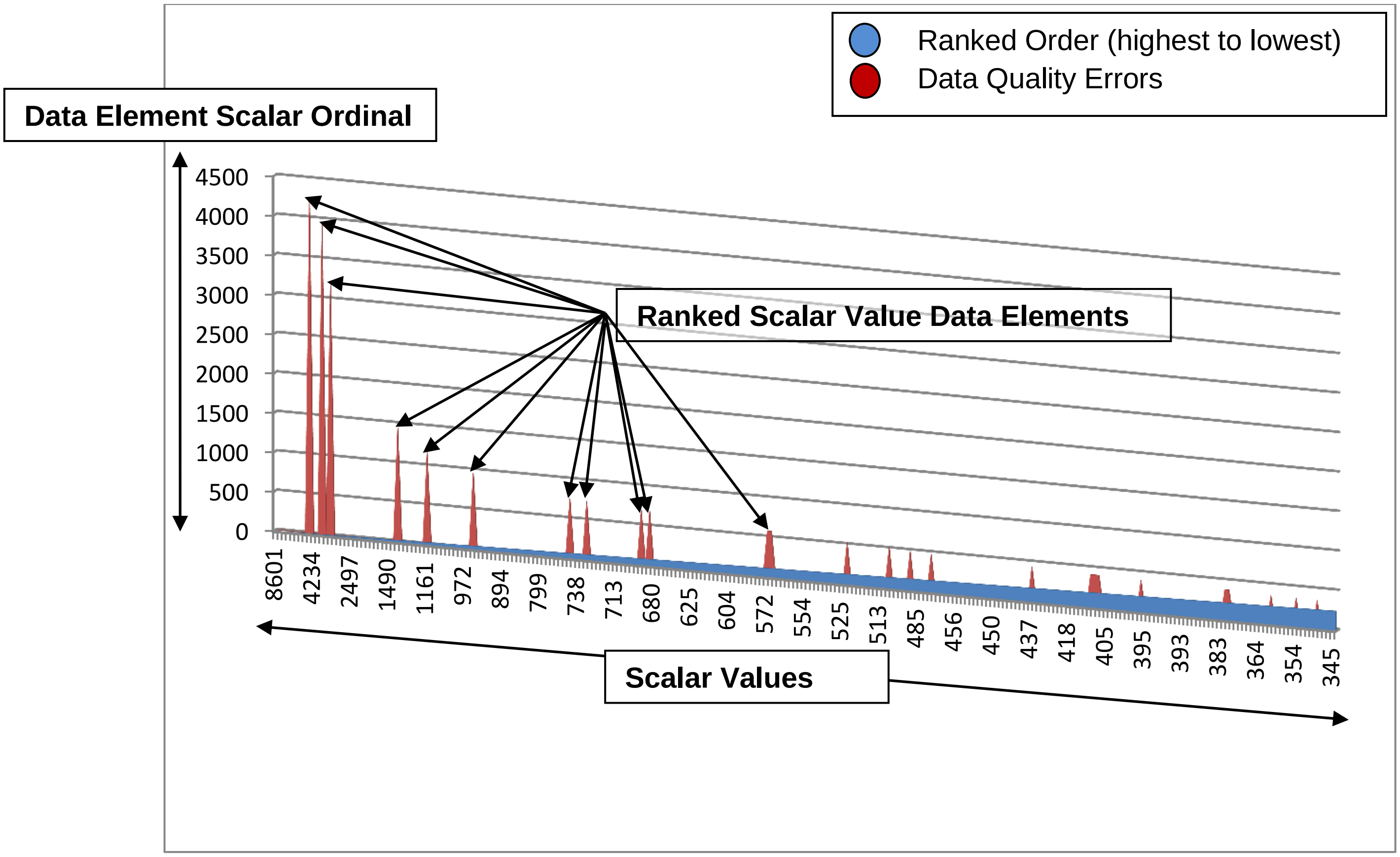


Figure 6-34 10% Simulated Data Quality errors

Figure 6-31 illustrates how a determination could be managed as a risk versus cost decision.

The higher the red spikes, the greater the ‘damage’ potential to an organisation.

The total scalar value of all contestable data elements is 730,336
The total contestable data elements are 17,397

The advantages of examining ranked data are still evident outside these boundaries, but they become less evident. If a database had errors approach 30% or more it would probably be unusable and so this or any other approach would probably not offer any distinct advantage.

7 CONCLUSION

The research question *"Is there a method that allows an organisation to identify data that presents the greatest quality assurance risk to an organisation?"* has been successfully addressed. The ranking approach to determining the relative importance of data and thus which data should be targeted for quality assurance has been developed and tested.

The results clearly demonstrate the usefulness of the ranking approach.

There would appear to be a best return area where the maximum benefit can be achieved from the ranking approach. This best return area would vary from database and organisational usage.

There may also be instances where data quality remediation cannot be conducted as the source of data may be uncontrolled or cannot be altered for evidential or legislative reasons. Ranking data elements and estimating percentage data quality errors against report usage and product lines offer significant decision making guides.

There is a significant advantage in terms of effectiveness using ranked data when searching for data quality errors in a data set. The number of data quality errors detected does not increase, but using the scalar value of contestable data elements, the organisational advantage is substantial.

This approach can also be used to determine if existing data quality detection approaches are successful.

The advantage appears also to improve when the percentage error rate increases past 1% data quality errors. Lower rates than this probably indicate that other data quality control measures are in place. Error rates exceeding 15% in a database would probably render the database unusable.

There would appear to be a best return area where the maximum benefit can be achieved from the ranking approach. This best return area would vary from database and organisational usage.

Using the North Wind Trader's database the best return area appears to be from 2% to 17% (depending upon random seeding strikes). The ranking benefit reduces as data quality data approach 25%.

The ranking tools allow an organisation to trial the sensitivity of their databases and determine the high sensitivity data sets that focus data quality analysis.

The next section describes the contributions that this research offers in the data quality data quality management domain.

7.1 LIMITATIONS AND ASSUMPTIONS

7.1.1 *LIMITATIONS*

Data quality errors may not be evenly distributed throughout a database, but may be clustered. This thesis assumes a random distribution to demonstrate a data ranking theory.

The models used in this paper describe the behavior of data usage in a *single* database. This set of experiments is presented as a proof-of-concept given that the North Winds Trader's database would, in a functioning organisation, be accompanied at least by a payroll, general journal, general ledger, creditor's ledger, debtor's ledger; assets register and inventory management data sets. The addition of these additional databases would change the scoring and ranking of the North Wind's Traders data sets. This database is publicly available so there are no ethical or commercial issues using this database as a test target.

This thesis assumes that all reports and data entry forms are used as intended. Often reports are designed and, as the business changes focus, they become redundant. There is no measurement of report usage frequency.

These experiments have been designed to suggest the concept that enables the viability to be demonstrated. It is expected that managing contestable data ranking across all databases in an organisation would offer a better picture of data rankings.

Using a ranking approach, this research aims to show that there is a repeatable approach to classifying and scoring data in an organisation.

A data element flagging facility allows simulated data quality errors to be introduced randomly across the database. Essentially each data element has its basic property list extended to include the accumulated ranking value for each data element as it is ranked through various data-related functions (such as queries, reports, and product lines and data entry screens).

The system allows each contestable data element to be randomly selected using a random number generator that selects a defined percentage of contestable data elements between element number 1 and *n* and then flag the data element as selected.

A means with which to identify a database's propensity (based on structure and usage) for data quality error propagation by tracking data elements as they are observed and measured in an organisational database.

7.1.2 *ASSUMPTIONS*

- Data quality measurement against a data element attracts the same level of cost and effort regardless of its significance in an organisation;

- Common data elements may be used by many levels within one organisation, often for different purposes and therefore attract different organisational values;

- The same data elements may attract different levels of significance across different organisations due to differences in business focus and internal processes; and

- An organisation's business directions often change over time thus potentially changing the focus on different data and data sets.

7.2 CONTRIBUTIONS

This thesis offers a range of contributions.

Firstly, addressing the problem and demonstrating that the problem has been addressed, makes a significant theoretical contribution.

Secondly, industry will benefit from an improved data quality examination model that minimises low-value data element testing as well as allowing a level of comfort as a result of the abbreviated examination method against high-value data.

Thirdly, the testing regime developed for the thesis is novel and provides the research community with an interesting research technique that can be used for further research in the data quality domain.

This thesis is an example of research of the "design research" type. As such it may be helpful for other researchers working in this new paradigm.

The literature review is very extensive and has been presented as a complete reference list for use by future researchers. This is a fifth contribution.

The design science approach used in this thesis (See 4.1, *Design Science* , p40) offers a structured model for artifact analysis, design and development.

The ranking tool offers a method by which like-named table columns can be collated against all tables in a database and the rules examined for consistency against organisational metadata standards.

A database can be profiled with varying data quality error rates to determine its potential impact in an organisation or at design time.

An organisation's databases could be profiled and the databases ranked in corporate value. This ranking, in turn, can then offer IS investment drivers as remedial action, security management or disaster recovery planning.

A database profile (shown as a graphed curve) will change as an organisation's business directions change. The rate of change offers an indicator that allows goodness-of-fit to be determined and potential database redesign or replacement considered.

Based on the potential to predict the percentage data quality errors based on a sample examination, a costed decision can then be devised to determine the level of data quality examination needed. A change in this profiling may indicate changed data entry or data quality management.

7.3 FUTURE RESEARCH DIRECTIONS

The research exposed a wide range of directions for future research. The literature review presented a range of knowledge gaps and opportunities for testing new ideas. In developing the software many design decisions were taken between plausible design options - these options remain to be explored.

Lastly, a range of contextual issues were revealed that also offer directions for further research.

1.1 Databases may present data quality errors in a non-random manner, making estimates of data quality error rates unlikely to be accurate. Ranking database in order of data values appears to remove error distribution bias. Further work testing the benefits of a ranked approach could contribute substantially to data quality management.
Contribution - Ranking data element by value could remove much of the data quality distribution bias, making data quality percentage estimates more accurate.

1.2 Development of a database design analysis tool that allows a mathematical model to be used to test and modify a database design to reflect expected propagation outcomes;
Contribution - to allow an analysis of database designs to optimize effectiveness and utility in an organisation.

1.3 Development of a statistical facility that allows unknown data quality errors to be estimated in a data population based on discoveries made when testing ranked data elements with some level of confidence.
Contribution - to allow an analysis of operational databases to predict, with a determined level of accuracy, what the overall data quality error rate might be. This approach, when combined with the ranking engine output, allows limited effort (and funding) to be applied to data quality management tactics.

1.4 Experiments across a complete organisational data set allowing all data (from all databases) used in the organisation to be ranked.
Contribution - to allow informed business decisions against the state of the database holdings and to inform IS investment decisions.

1.5 Experiments can profile databases using 'goodness-of-fit" in an organisation thus allowing informed replacement or correction decisions to be formed .
Contribution - to allow informed business decisions against the state of the database holdings and to inform IS investment decisions.

1.6 Mapping an organization against data holdings by value can be used to inform disaster recovery planning and priorities as a business and IT alignment measure.
Contribution would be to justify IS expenditure against high value IS targets.

REFERENCES

Abate, M. L., Diegert, K. V., Allen, S.N.L. and Allen,H. (1998)
A Hierarchical Approach to Improving Data Quality.
Data Quality, Sept 1998

Abderdeen Group (Harte-Hanks), (2007)
The Cost of Quality : Benchmarking Enterprise Quality Management
viewed 29 July 2011
http://www.aberdeen.com/aberdeen-library/4121/RA-quality-matters.aspx

Australian Competition and Consumer Commission (ACCC) (2010)
Review of the Australian product safety recalls system
Last viewed 1 July 2011
http://www.accc.gov.au/content/index.phtml/itemId/930113

Australian Competition and Consumer Commission (ACCC) (2011)
Motor vehicle recall samples
Last viewed 18 July 2011

BMW Recalls http://www.recalls.gov.au/content/index.phtml/itemId/952885
Honda Recalls http://www.recalls.gov.au/content/index.phtml/itemId/952862
Porsche Recalls http://www.recalls.gov.au/content/index.phtml/itemId/952901
Toyota Recalls http://www.recalls.gov.au/content/index.phtml/itemId/952857
Volvo Recalls http://www.recalls.gov.au/content/index.phtml/itemId/952882

Alstyne, Marshall W.V. (1999)
A proposal for valuing information and instrumental goods.
University of Michigan, USA
ICIS '99 Proceedings of the 20th international conference on Information Systems pp328-345

Aspect Computing (2002)
(no longer exists), purchased by KAZ Computer Services
(no longer exists), purchased by Fujitsu Australia
viewed 24 February 2012
http://www.fujitsu.com/au/

Australia Competition and Consumer Commission (ACCC) (2010)
Product recall notices
http://www.recalls.gov.au/content/index.phtml/itemId/952864
viewed 25 June 2011

Auditor General, Australian National Audit Office (1998-1999)
ATO Data and Systems Quality – An ANAO Discussion Paper

Reference to:
The Management of Tax File Numbers in The Australian Taxation Office
The Auditor-General Audit. Report No 37 1998-99, p84

Auditor General, Australian National Audit Office (2003-2004)
ATO Data and Systems Quality – An ANAO Discussion Paper
Reference to:
HIC ANNUAL REPORT 2003–04 INTERNAL CONTROL FRAMEWORK

Badri, Masood A., Davis,Donald & Davis,Donna (1995)
A study of measuring the critical factors of quality management
International Journal of Quality & Reliability Management, Vol.12, Number 2

Baskerville, R., Pries-Heje, J., & Venable, J. R. (2008)
Strategies for Design Science Research Evaluation.
European Conference on Information Systems (ECIS) at AIS Electronic Library (AISeL), Paper 87, Folio 13

Baskerville, R., Pries-Heje, J., & Venable, J.R. (2009).
Soft design science methodology.
Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology - DESRIST '09, 1
New York, New York, USA: ACM Press

Bartneck, C. (2009)
Define Design Science.
Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology

Batini, Carlo, Cappiello, Cinzia, Francalanci, Chiara & Maurino, Andrea (2009)
Methodologies for Data Quality Assessment and Improvement
ACM Computing Surveys, 41(3), 1-52, Article 16

Belasco, Amy, (2011)
The Cost of Iraq, Afghanistan, and Other Global War on Terror Operations Since 9/11
Congressional research Service, 7-5700 RL33110

Bollen, J., Alamos, L., Rodriguez, M. A., & Sompel, H. V. D. (2007)
MESUR : usage-based metrics of scholarly impact
Science. viewed April 2 2011
http://www.mesur.org/MESUR.html

Bowen, P.P.L., Jarke,M., Aachen, R. & Madnick,S.E. (1999)
Panel Data Quality In Internet Time , Space , and Communities Chair
Journal of Information Systems, 713-716.

Bowtell, Matthew (1999)

Dimensions Of Information Systems Success.
Information Systems, 2 November 1999.

Bradley (1968)

Distribution-Free Statistical Tests
Chapter 12
viewed 16 February 2010
http://itl.nist.gov/div898/handbook/eda/section3/eda35d.htm

Bryant, R. E., & Digney, J. (2007)

Data-Intensive Supercomputing: The Case for DISC
Forbes: Fall 2007, pp1-10
Parallel Data Laboratory, Carnegie-Mellon University, USA

Burgess, M. (2007)

Data Quality – What is it , and do we agree ?
Lecture slides 15th June 2007
School of Computer Science, Cardiff University UK

Canadian Institute for Health Information (CIHI) (2005)

CIHI 61 June 2005 ed.
viewed 5 December 2009
http://www.cihi.ca/cihiweb/en/downloads/Data_Quality_Framework_2004_e.pdf

Calero, C., & Piattini, M. (n.d.)

A Data Quality Measurement Information Model Based On ISO/IEC 15939
Indra-UCLM Research and Development Institute
Ronda de Toledo s/n – 13003 Ciudad Real, Spain

Carlos, A., & Maçada,G.

A Model for Information Quality in the Banking Industry – the Case of the Public Banks in Brazil. Analysis.

Chapman, A. D., & Speers, L. (1991)

Principles Of Data Quality
Global Biodiversity, (Chrisman)

Chaudhuri, S., Das, G., Hristidis, V., & Weikum, G. (2006).

Probabilistic Information Retrieval Approach for Ranking of Database Query Results
University of Texas at Arlington. *Database*, *31*(3), 1134-1168.

Cleven, A., Gubler, P., & Hüner, K. M. (2009).

Design alternatives for the evaluation of design science research artifacts.

Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology - DESRIST ’09, 1. New York, New York, USA: ACM Press

Columbia University Technical Report CUCS-028-06 (n.d.)
Practical Preference Relations for Large Data Sets
Viewed December 2009,
http://app.cul.columbia.edu:8080/ac/bitstream/10022/AC:P:29465/1/411.pdf

Cong, G., Fan, W., Geerts, F., Jia, X., & Ma, S. (2007)
Improving Data Quality : Consistency and Accuracy.
Microsoft Research, Bell Laboratories, University Edinburgh, and Hasselt University

Dasu, T., Vesonder, G. T., & Wright, J. R. (2003)
Data quality through knowledge engineering.
Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’03, 705. New York, New York, USA: ACM Press

DataFlux Corporation LLC (2007)
The Cost of Quality
©DataFlux Corporation LLC, Cary, USA
viewed 24 July 2011
http://www.dataqualitypro.com/storage/dataflux/WP038%20The%20Data%20Governance%20Maturity%20Model.pdf

DataFlux White Paper, (2010)
The Data Governance Maturity Model.
Integration - The VLSI Journal, 44(0).
viewed 20 July 2011
http://www.dataqualitypro.com/storage/dataflux/WP038%20The%20Data%20Governance%20Maturity%20Model.pdf

de Corbière, Francois, (2009)
Data Quality and Interorganizational Information Systems: The Role of Electronic Catalogues
AMCIS 2009 Proceedings. Paper 125.
Viewed http://aisel.aisnet.org/amcis2009/125

DeLone, W. H., & McLean, E. R. (2003)
The DeLone and McLean Model of Information Systems Success : A Ten-Year Update.
Journal of Management Information Systems, *19*(4), 9-30.

Dishaw,M. T., Strong,D.M., & Bandy,D.B. (2002)
Extending The Task-Technology Fit Model With Self-Efficacy Constructs.
Information Systems, pp1021-1027.

Dasu,T., Vesonder, G.T. & Wright, J.R. (2003)

Data quality through knowledge engineering.
Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’03.
viewed 20 November 2010
http://portal.acm.org/citation.cfm?doid=956750.956844.

Deming, W. E., & Shewhart, W. (2008)

Japanese Manufacturing “ Out of the Crisis ”
** Draft 08-15 **. *New York*, 15-17.pp23-24
AHP Lean Sigma Op Research\Deming and Toyota from 8/16/08

Department of Communications, Information Technology & the Arts Information Quality & ICT (2006)

Exploratory Research Report : Final

Dunne,P.E.,, Parsons,S., Wooldridge,M., Mcburney,P., Dunne,E., & Hunter,A. (2009)

Sierra and Castelfranchi (ed.)
Inconsistency Tolerance in Weighted Argument Systems
Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009), Decker, Sichman, , May, 10–15, 2009, Budapest, Hungary, 851-858.

Embury, S. M., Missier, P., Sampaio, S., & Greenwood, R. M. (2009)

Incorporating Domain-Specific Information Quality Constraints into Database Queries
Quality, 1(2), 1-31

Eckerson, W. W. (2002)

Data Quality and the Bottom Line
TDWI Report Series
THE DATA WAREHOUSING INSTITUTE
viewed 12 June 2011
www.dw-institute.com

English,L.P. Wells,David, Richardson,Sid, Mimno,Pieter (u.d.)

Barton, Louis (Ed.),
TDWI Report Series (2nd ed., p. 36).
Seattle, WA 98188: The Data Warehousing Institute.

English, L.P. (2011)

Applying Deming’s 14 Points To Information Quality : The 14 Points of Total Information Quality Management
© 2011 INFORMATION IMPACT International , Inc . *Business*, 1-14.

Eppler, M.J. & Wittig, D. (2000)

Conceptualizing Information Quality: A Review of Information Quality Frameworks from the Last Ten Years Goals of an Information Quality Framework.
Proceedings of the 2000 Conference on Information Quality

Etezadi-Amoli, J. and A. F. Farhoomand (1996)

A Structural Model of End User Computing Satisfaction and User Performance
Information & Management, 30, (1996), 65-73.

Even, A., & Shankaranarayanan, G. (2007)

Utility-Driven Assessment of Data Quality.
Data Base For Advances In Information Systems, 38(2), p75-93

Feng, Y., Agrawal, D., El Abbadi, A., & Singh, A. (2005)

Scalable ranking for preference queries.
Proceedings of the 14th ACM international conference on Information and knowledge management - CIKM ’05, 0(Figure 1), 313. New York, New York, USA: ACM Press

Fogarty, G.J., Armstrong, D.B., Dimbleby, J. & Dingsdag (2003)

Exploring user satisfaction with information systems in a regional small business context
ed. J. Ang & S. Knight
Proceedings of Delivering IT and e-business value in networked environments:
14th Australasian Conference on Information Systems, Perth, WA, 26-28 November, WeB Centre, Edith Cowan University, Joondalup, WA, pp. 1-9.

Frontier Software Pty Limited

Viewed 1st December 2010
http://www.frontiersoftware.com/

Gelderman, M. (1998)

Translation and validation of the Doll and Torkzadeh end-user computing satisfaction instrument.
Proceedings of the Thirty-First Hawaii International Conference on System Sciences, *6*(c), 537-546. IEEE Comput. Soc.

Gershon, Peter (2008)

Review of the Australian Government’s use of Information and Communications Technology
viewed 5 December 2009
http://www.ag.gov.au/cca

Goasdoué,V., Duquennoy,D., Laboisse,B., & Sylvaine,N. (n.d.)

A Comparison Framework for Data Quality Tools.

Goodhue, D. L., & Thompson, R. L. (1995)

Task-Technology Fit and Individual Performance
MIS Quarterly, *19*(2), 213. doi:10.2307/249689

Guimaraes, T. & Igbaria, M., (1996)

Assessing User Computing Effectiveness : An Integrated Model.

Journal of Organizational and End User Computing, 9(2), pp.1996-1998.

Hamlet, R. (1983)
Random Testing, (1), 1-25.
viewed 10 June 2011
citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.77.5425&rep

Hasan,S., & Padman,R. (2006)
Analyzing the effect of data quality on the accuracy of clinical decision support systems: a computer simulation approach.
AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium, 324-8.
viewed 10 June 2011
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1839724&tool=pmcentrez&rendertype=abstract

Heeren, C., & Pitt, L. (2005)
Maximal Boasting,
August 21-(KDD'05, August 21–24, 2005), p580-585.

Haveliwala, T. H. (2003)
Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search.
IEEE Transactions on Knowledge and Data Engineering, *15*(4), 784-796. Vol. 15, NO. 4, July/August 2003

Hevner, Alan R, Salvatore,T.March, Park,Jinsoo, Ram,Sudha, (2004)
Design Science in Information Systems Research
Published MIS Quarterly, Oct. 2004
viewed 5 December
http://piom.uni-graz.at/pdf/MIS%20Quarterly%2028,%202004_Design%20Science%20in%20Information%20Systems%20Research.pdf

House of Commons Foreign Affairs (2002)
The Decision to go to War in Iraq
Ninth Report of Session 2002–03

Huff, R. A., Guynes, C. S., & Golladay, R. M. (1995)
Ethics, Information Systems, and the Information Highway
University of North Texas
Computers and Society, (March, 1995).

Iivari, Juhani (2004)
An Empirical Test of the DeLone-McLean Model of Information System Success
ACM Digital Library
Volume 36, Issue 2 (Spring 2005), pp8-27
viewed 5 December 2009

http://portal.acm.org/citation.cfm?id=1066149.1066152&coll=portal&dl=ACM&CFID=1181471&CFTOKEN=40269323

Ilyas, I. F., & Das, G. (2007)
Report on the First International Workshop on Ranking in Databases (DBRank'07)
ACM SIGMOD Record, *36*(4), 49

Juran, J. & Godfrey, A.B. (1999)
Juran's Quality Handbook. 5th ed.
McGraw-Hill, New York.

Jurison, J. (1999)
Software Project Management : The Manager's View
Communications, *2,* (September 1999)

Kalashnikov,D.V. & Mehrotra, S. (2006)
Domain-independent data cleaning via analysis of entity-relationship graph. ACM Transactions on Database Systems, 31(2), 716-767

Kaluzny, A.D. (1982)
Quality assurance as a managerial innovation: a research perspective. Health services research, 17(3), 253-68.
Viewed 22 May 2011
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1068680&tool=pmcentrez&rendertype=abstract.

Knight, Shirlee-ann & Burn,Janice (2005)
Developing a Framework for Assessing Information Quality on the World Wide Web Introduction – The Big Picture What Is Information Quality ? *Science Journal*, *8*.

Kovac, R., Lee, Y. W., & Pipino, L. L. (n.d.)
Total Data Quality Management : The Case of IRI.
Information resources Group, Cambridge Research Group and University of Massachusetts

Krogstie, J. (1996)
A Semiotic Approach to Quality in Requirements Specifications.
Quality, Chapter 14, 1-19

Kuechler, B., & Vaishnavi, V. (2011)
Promoting Relevance in IS Research: An Informing System for Design Science Research Rigor and Relevance: A Recurring Dilemma

Langville, Amy N. & Meyer, Carl D., (2006)
Google's PageRank and Beyond :The Science of Search Engine Rankings

Princeton University Press, Princeton University, USE

Lee, Chia-Jung, Chen, Ruey-Cheng, Kao, Ruey-Cheng, Cheng, Pu-Jen (2009)
A term dependency-based approach for query terms ranking.
Proceeding of the 18th ACM conference on Information and knowledge management - CIKM 2009, p1267.
viewed 23 January 2011
http://portal.acm.org/citation.cfm?doid=1645953.1646114

Lei, Y., Uren, V., & Motta, E. (2007)
A framework for evaluating semantic metadata
Proceedings of the 4th international conference on Knowledge capture - K-CAP '07, 135.
New York, New York, USA: ACM Press

Levis, M., Helfert, M., & Brady, M. (n.d.)
Information Quality Management : Review Of An Evolving Research Area
Research Paper Information Quality(IQ)
Retrieved 01/10/2010
http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.84.1748

Liu, E. Rong, Kumar, Akhil, & Aalst, Wil van der. (2007)
A Formal Modeling Approach for Supply Chain Event Management
Decision Support Systems, Volume 43, Issue 3
Integrated Decision Support, April 2007, Pages 761-778

Liu, Y., Zhang,L., Song,R., Nie,J.Y., & Wen,J.R. (2009)
Clustering queries for better document ranking.
Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09, 1569. New York, New York, USA: ACM Press

Madnick S.E. & Wang R.Y. (1992)
Introduction to total data quality management (TDQM) research program.
TDQM-92-01, Total Data Quality Management Program,
MIT Sloan School of Management.

Madnick S.E., Wang,R.Y., Dravis,F., & Chen,X. (2003)
Improving the Quality of Corporate Household Data: Current Practices and Research Directions
SSRN Electronic Journal, 92-104

Madnick, S.E., Wang,R.Y., Lee,Y.W. & Zhu,H. (2009)
Overview and framework for data and information quality research.
ACM J. Data Inform. Quality 1, 1, Article 2 (June 2009) 22 pages.

Marcey L. Abate, Kathleen V. Diegart, Heather W.Allen (1998)
A Hierarchical Approach to Improving Data Quality

COMMUNICATIONS OF THE ACM
February 1998/Vol. 4, No. 1, viewed 22 January 2010
http://portal.acm.org/citation.cfm?id=269012.269021&jmp=cit&coll=ACM&dl=ACM&CFID=64034208&CFTOKEN=71793119#

McKnight, William (2010)
Data Quality for the Next Decade
Information Management Magazine, Nov/Dec 2010

Microsoft™ 2007
OneNote ©
viewed May 5 2011
http://office.microsoft.com/en-au/onenote/

Mill, J. S. (1846)
A system of logic, ratiocinative and inductive, being a connected view of the principles of evidence and the methods of scientific investigation. 8th Ed.
Methods. New York: Harper & Brothers, Franklin Square.
Electronically Published by Project Guttenberg
Viewed 15 July 2009
http://www.gutenberg.org/ebooks/27942

Missier, Paolo, Embury,Suzanne, Greenwood,Mark, Preece,Alun, Jin,Binling (2006)
Capturing and Exploiting the User Perspective on Data Quality
VLDB '06, September 1215, 2006, Seoul, Korea,
viewed 5 December 2009
http://www.springerlink.com/content/q78x31366106t7x5/fulltext.pdf?page=1

Moore, Susan, (2007)
Gartner Inc., Newsroom
Viewed 12 June 2011
http://na2.www.gartner.com/it/products/newsroom/index.jsp

Needham,D.M., Sinopoli,D.J., Dinglas,V.D., Berenholtz,S.M., Korupolu,R., Watson,S.R. (2009)
Improving data quality control in quality improvement projects.
International Journal for Quality in Health Care *and*
Journal of the International Society for Quality in Health Care / ISQua, 21(2), 145-50

O'Brien, T. (2011)
Poor Data Can Cost You Money and Get You Sued.
International Association for Information and Data Quality

Offermann,P., Levina,O., Schönherr,M., & Bub,U. (2009).
Outline of a design science research process.

Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology - DESRIST ’09, 1. New York, New York, USA: ACM Press.

Otto,B., & Ebner,V. (2010)
Measuring Master Data Quality Findings from an Expert Survey
Information Systems, 1101-1112

Page,L., Brin,S., Motwani,R., & Winograd,T. (1998)
The PageRank Citation Ranking: Bringing Order to the Web,
Stanford Digital Libraries Working Paper, 1998

Parker,M., Stofberg,C., Harpe,R.D., Venter,I., & Wills,G. (2006)
DATA QUALITY: HOW The Flow Of Data Influences Data Quality In A Small To Medium Medical Practice.
Community informatics for developing countries, 31 August -2 September 2006
Cape Town, South Africa.

Parssian,A., Sarkar,S., & Jacob,V.S. (2004)
Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product. Management Science, 50(7), 967-982

Peffers, Ken, (2011)
Targeted Service Co-Design Theory
MIS Quarterly, (February).

Peffers, K., Tuunanen, T., Gengler, C. E., Rossi, M., Hui, W., Virtanen, V., & Bragge, J. (2006)
The Design Science Research Process: A Model For Producing And Presenting Information Systems Research.
DESRIST,Claremont, CA., USA

Peffers,K., Tuunanen,T., Rothenberger,M.A., & Chatterjee,S. (2008)
A Design Science Research Methodology for Information Systems Research
Journal of Management Information Systems
Vol. 24 No. 3, Winter 2008, viewed 5 August 2009
http://www.jmis-web.org/articles/v24_n3_p45/index.html

Petter,S., & Murphy,J.D. (2010).
A Design Science Based Evaluation Framework for Patterns.
Data Base for Advances In Information Systems, 41(3)

Pipino,L.L., Lee,Y.W., & Wang,R.Y. (2002)
Data quality assessment
Communications of the ACM, 45(4ve), pps211-218

Piprani,B., Ernst,D., & Canada,S. (n.d.)

A Model for Data Quality Assessment. Management

Pham,Tu (2007)

Australian Capital Territory, Auditor General's Report
ACT Auditor General
ACT AG Report May, 2008 p.8, Section 1.6
ACT AG Report May, 2008 p.10, Section 1.8

Pon,R.K., & Cárdenas, A.F. (2005)

Data quality inference.
Proceedings of the 2nd international workshop on Information quality in information systems - IQIS '05, 105. New York, New York, USA: ACM Press

Pitney-Bowes Business Insight Data Flow (u.d.)

Results from the Information Difference survey document initiatives and the state of data quality today;
Lanham, MD 20706-1844, USA, v**iewed 5 December 2009**
http://www.information-management.com/issues/2007_56/10015032-1.html

Pitney-Bowes (2009)

The Sad State of Data Quality, viewed 5 December 2009
http://www.information-management.com/issues/19_8/the-sad-state-of-data-quality-10016576-1.html

Price,R., & Shanks,G. (2005) (a)

Empirical Refinement of a Semiotic Information Quality Framework.
Proceedings of the 38th Annual Hawaii International Conference on System Sciences, 00(C), 216a-216a

Price,R., & Shanks,G. (2005) (b)

A semiotic information quality framework: development and comparative analysis.
Journal of Information Technology, 20(2), 88-102

Price,R., Neiger,D., & Shanks,G. (2008)

Developing a Measurement Instrument for Subjective Aspects of Information Quality
Quality, *22,*January, pp49-74.

Pries-Heje,Jan, Baskerville,Richard, & Venable,John R. (2008)

Strategies for Design Science Research Evaluation
ECIS 2008 Proceedings. Paper 87.

Raden ,Neil (2006)

The Growing Externalization of Business Processes Requires the Discovery of Meaning in Data
Hired Brains Research Inc, Santa Barbara, CA 93105, USA

(Sponsored by Silver Creek Systems, Inc., Westminster, CO 80021, USA)

Redman,Thomas C. (1998)
The Impact of Poor Data Quality on the Typical Enterprise
Communications of The ACM
February 1998/Vol. 41, No. 2, pp.79-81, viewed 5th December 2009
http://portal.acm.org/citation.cfm?id=269012.269025&coll=ACM&dl=ACM&CFID=64034208&CFTOKEN=71793119

Ross, Kenneth A., Stuckey,Peter J., Marian, Am´elie (2007)
Practical Preference Relations for Large Data Sets
Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop
Viewed 5th December 2009,
http://www.computer.org/portal/web/csdl/doi/10.1109/ICDEW.2007.4400997

Rudra,A., & Yeo,E., (1999)
Key issues in achieving data quality and consistency in data warehousing among large organisations in Australia.
Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999, viewed 30 November 2010
http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=772757

Sarbanes-Oxley Act of 2002 (2002)
The One Hundred and Seventh Congress of the United States of America (2nd Session) viewed 5 December 2009
http://fl1.findlaw.com/news.findlaw.com/hdocs/docs/gwbush/sarbanesoxley072302.pdf

Scannapieco,M., Bertino,E., & Elmagarmid,A. (2007)
Schema and Data Matching
Elements, pp653-664.

Scannapieco,M., Mecella,M., Catarci,T., Cappiello,C., Pernici,B., Mazzoleni,F. (2001)
Comparative Analysis of the Proposed Methodologies for Measuring and Improving Data Quality and Description of an Integrated Proposal

Seddon, Peter B., Staples,D.Sandy, Patnayakuni,Ravi, Bowtell,Matthew J. (1996)
The IS effectiveness matrix: the importance of stakeholder and system in measuring IS Success
Association for Information Systems (via *University of Melbourne, Australia*)
ICIS '98 Proceedings of the International Conference on Information Systems, *p165-176*
December 1998, ICIS ’98

Seddon,Peter.B., Staples,D.Sandy., & Patnayakuni,Ravi. (1999)
Dimensions of information systems success.
Communications of AIS

Volume 2, Article 20 (November, 1999)

Skinner, Erik (2009)
Californian Community Colleges Chancellor's Report, Oct 23rd,
Vice Chancellor's Financial Services
Viewed 5th December 2009
http://www.faccc.org/images/ARRA_Funding_Guidelines_for_Districts.pdf

Standard and Poor's 2004 Review : Global Index Review (2004)
Clients Services, New York
pp4-20, 1999 – 2004 comparisons
p2 S&P Global 1200, 1999 – 2004 comparisons

Stvilia, B. (2008)
A Workbench for Information Quality Evaluation, *58*(2007)
Florida State University, Tallahassee, Florida, USA

Stiglitz, B. J. E., & Bilmes, L. J. (2010)
The Washington Post September 5, 2010
The true cost of the Iraq war : $3 trillion and beyond

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997)
Data Quality in Context.
Communications of the ACM, *40*(5), 103-110.

Supply Chain Council, (2009)
Supply Chain Operations Reference (SCOR®)
Viewed 23 March 2011
http://supply-chain.org/ and
http://supply-chain.org/about/history

Igbaria,Magid & Tan,Margaret (1998)
The Virtual Workplace
Idea Group Publishing, Covent Garden, London, UK

Tayi, G. K., & Ballou, D. P. (1998)
Examining Data.
Communications of the ACM, *41*(2), pp54-57

Telang, A., Mishra,R. & Chakravarthy,Sharma (2007)
Ranking Issues for Information Integration
First International Workshop on Ranking in Databases in Conjunction with ICDE
Viewed 5 December 2009,
http://portal.acm.org/citation.cfm?id=1361348.1361361&coll=ACM&dl=ACM&CFID=64034208&CFTOKEN=71793119

Teng, J. T. C., Calhoun, K. J., Raeburn, S., & Wong, W. (1984)

Is The East Really Different From The West : A Cross-Cultural Study On Information Technology And Decision Making
University of Ulsan, Korea
United Kingdom. *Decision Sciences*, (1984), 40-46.

Teo, T., Wong, P. K., & Ee Hui Chia. (2000)

Information technology (IT) investment and the role of a firm: an exploratory study
International Journal of Information Management, *20*(4), 269-286

Powell, James (2011)

The Data Warehouse Institute (TDWI)
BI Not Meeting the Needs of Marketers and Merchandisers
retrieved 12 November 20011
http://tdwi.org/articles/2011/10/04/bi-not-meeting-needs.aspx

The National Center for Manufacturing Sciences & University of Michigan (2002)

University of Michigan and Tauber Manufacturing Institute (2002)
COST OF INFORMATION ASSURANCE, (August), 1-26
Ann Arbor, Michigan
viewed 21 August 2010
http://www.ncms.org/

Thomson Reuters (u.d.)

EndNote X4™,
viewed May 5 2011
http://www.endnote.com/eninfo.asp

Timmins, N. (2007)

Financial Times 13th July 2007.
Financial Times (UK)
Viewed 2 December 2010
http://www.ft.com/home/uk

Torkzadeh, G. & Doll, W.J. (1987)

The relationship of MIS steering committees to size of firm and formalization of MIS planning. *Communications of the ACM*, 30(11), pp.972-978.
viewed 19 March 2011
http://portal.acm.org/citation.cfm?doid=32206.32213

Waddington, David (2009)

The Sad State of Data Quality
Information Management Magazine
November 2009

Wagner,Stefan & Meisinger,Michael (2006)

Integrating a model of analytical quality assurance into the V-Modell XT
Proceedings of the 3rd international workshop on Software quality assurance - SOQUA
Viewed 5 December 2009
http://portal.acm.org/citation.cfm?doid=1188895.1188906.

Wang,Richard & Strong,Diane M. (1996)
Beyond accuracy - What data quality means to data consumers.
Journal of Management Information Systems
Armonk: Spring 1996.

Wand,Y., & Wang,R.Y. (1996)
Anchoring Data Quality Dimensions in Ontological Foundations.
Communications of the ACM, 39(11), 86-95.

Wang,R.Y. (1999)
Total Data Quality Management.
Communications of the ACM, 41(2).

Watson,Paul, Tayi,Giri Kumar & Ballou,Donald P. (1998)
Examining Data Quality
COMMUNICATIONS OF THE ACM
February 1998/Vol. 41, No.2 pp54-57
Viewed 5 December 2009
http://portal.acm.org/citation.cfm?id=269012.269021&jmp=cit&coll=ACM&dl=ACM&CFID=64034208&CFTOKEN=71793119#

Watts,S., Shankaranarayanan,G., & Even,A, (2009)
Data quality assessment in context: A cognitive perspective.
Decision Support Systems, 48(1), 202-211

Welzer, T., Brumen, B., Golob, I., & Druovec, M. (2002)
Medical diagnostic and data quality.
Proceedings of 15th IEEE Symposium on Computer-Based Medical Systems (CBMS 2002), (CBMS), 97-101. IEEE Computer. Society.

Willshire,Mary-Jane & Meyen,Donna (1997)
A Process for Improving Data Quality
Data Quality - September, Vol.3 No 1
viewed 5 December 2009
http://www.dataquality.com/997meyen.htm#introduction

Wixom, B. H., & Watson, H. J. (2001)
An Empirical Investigation Of The Factors Affecting Data Warehousing Success
MIS Quarterly, *25*(1), 17-41

Yakout,Mohamed, Elmagarmid,Ahmed K., Neville,Jennifer (2010)

Ranking for Data Repairs
4th DBRank workshop, ICDE 2010 , Long Beach, California, USA.
Viewed 24 August 2010
http://www.cs.purdue.edu/homes/myakout/

Zhou,Dengyong, Weston,Jason, Gretton,Arthur, Bousquet,Olivier & Schlkopf,Bernhard (2003)
Ranking on Data Manifolds
Max Planck Institute for Biological Cybernetics, 72076 Tuebingen, Germany
viewed 5 December 2009
http://www.kyb.mpg.de/bs/index.html

APPENDICES

APPENDIX A - RANDOM SEEDING TRIALS

Introduction

Each percentage rate for flagged errors was introduced using a random number generator that is triggered with a new seed comprising the seconds between current date and time to seconds since to today's data and time. Therefore each random run (expectedly) generates a slightly different pattern.

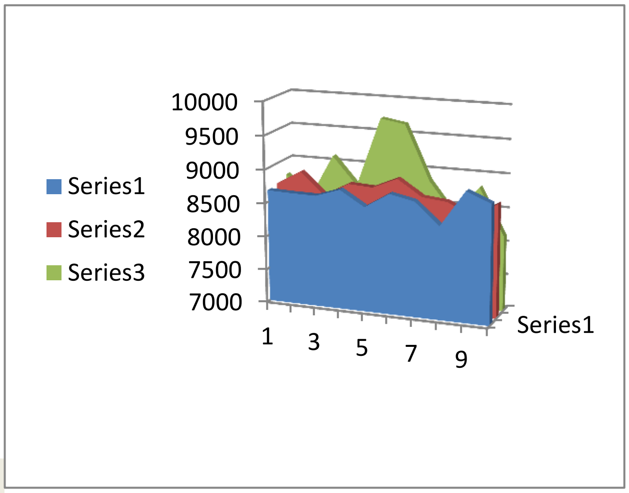
The following scripts were generated automatically by the Ranking Software as a control facility over 10 trials for each 1%, 5% and 10%.

Each percentage random error set performs some basic statistics again the flagged data elements and the data element population as a whole. The aim of this information is to show an even distribution is illustrated by showing the numeric midpoint for data elements (always 8698) and that the median of the randomly flagged data elements is similar. This numeric outcome is verifying by observing the random number pattern generated on the screen when the random set has been created.

The intention of the random generation sequence was to present the flagged data elements so that the "*statistical independence among test points allows statistical prediction of significance in the observed results*", Hamlet,1983,p3

The outcomes have been graphed showing the independence each run for each percentage. Note that the sample size is 10, so outliers may be more prominent that with larger samples. An assumption is that data quality errors may not always be random.

	Series 1	**Series 2**	**Series 3**
Trial#	**10%**	**5%**	**1%**
Trial 1	8677	8700	8764
Trial 2	8665	8909	8427
Trial 3	8660	8587	**9099**
Trial 4	8781	8787	8685
Trial 5	8559	8766	**9712**
Trial 6	8785	8927	**9650**
Trial 7	8709	8689	8845
Trial 8	8385	8652	8375
Trial 9	8923	8477	8815
Trial 10	8773	8693	8120
Average	**8691.7**	**8718.7**	**8849.2**
St Dev.	**145.5**	**136.8**	**518.3**



Note, the series 3 appears to have 3 outlier data components (highlighted in bold and underline). The effect on the average is minor and considering the nature of data quality errors, it was decided to allow the occasional outlier to persist. These figures are the mean of several random trial runs.

1% Random population Trials and outcomes

Trial 1

Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 1% tag rate. Data element midpoint = 8698. Tagged data elements midpoint=8764

Trial 2

Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 1% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8427

Trial 3

Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 1% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 9088

Trial 4

Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 1% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8565

Trial 5

Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 1% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 9712

Trial 6
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 1% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 9650

Trial 7
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 1% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8845

Trial 8
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 1% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8375

Trial 9
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 1% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8815

Trial 10
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 1% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8120

5% Random population Trials and outcomes

Trial 1
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 5% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8700

Trial 2
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 5% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8909

Trial 3
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 5% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8587

Trial 4
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 5% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8787

Trial 5
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 5% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8766

Trial 6
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 5% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8927

Trial 7
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 5% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8689

Trial 8
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 5% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8652

Trial 9

Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 5% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8477

Trial 10

Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 5% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8693

10% Random population Trials and outcomes

Trial 1
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 10% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8677

Trial 2
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 10% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8665

Trial 3
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 10% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8669

Trial 4
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 10% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8781

Trial 5
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 10% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8559

Trial 6
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 10% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8785

Trial 7
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 10% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8709

Trial 8
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 10% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8535

Trial 9
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 10% tag rate. Data element midpoint = 8698. Tagged data elements midpoint = 8923

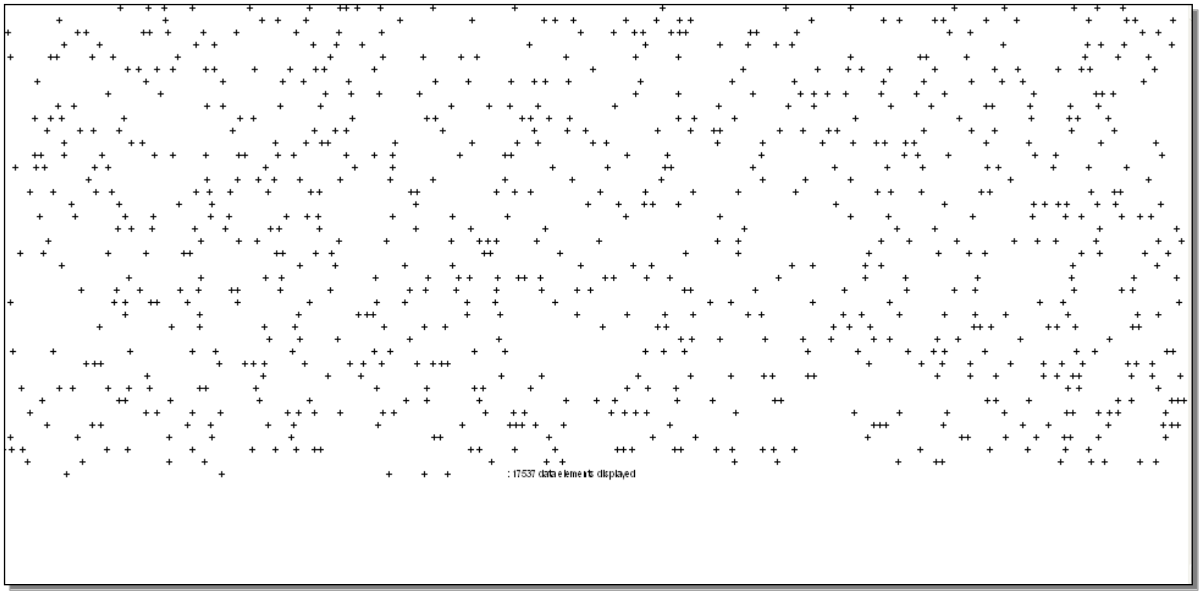
Trial 10
Database: Northwind.mdb. Contains 17525 contestable data elements. These elements have been tagged representing a 10% tag rate. Data element midpoint = 8698.   Tagged data elements midpoint = 8773

Random Error Seeding

A mechanism has been designed to select and assign a selected percentage flagging all contestable data elements across all the tables in the database. For example, a 5% flagging rate selects 854 data elements from 17525 data elements.

For a random selection of the all contestable data elements, the randomness of the flagging mechanism by writing a '+" at the random number value in a string the length of the 'n' (in Northwind Traders database, there are 17525 data elements):

For example, the 17525 contestable data elements are represented in a single string that has '+' for flagged and " " for not flagged. This string is displayed from left to right and wrapping at the right boundary to show the random distribution of flagged data elements.



Repeated tests at 5% error rate were executed to demonstrate random data flagging and therefore random data quality error simulation. Note that although each generation of random contestable data elements offers a similar spread pattern, the random number generator is seeded with today's date and so generates different patterns each run.

Random generation for the data quality error simulator. This facility allows any percentage error introduction from .1% through to 99.9% and, once completed illustrates the distribution of the contestable data element error flags.

APPENDIX B - COMMERCIAL DATA QUALITY PROVIDERS

Information Difference
http://www.informationdifference.com/dq_landscape.html, *(Last accessed 1 December 2009)*

These vendors provide data quality issue detection using a variety of automated means. They do not, however, assess data quality errors based on each user in an organisation, but rather a uniform approach.

Vendor	Brief Description	Website
AMB New Generation Data Empowerment	Chicago-based vendor specializing in just-in-time data quality with in-stream profiling and outlier detection.	www.ambpdm.com
Address Doctor	Vendor that specializes in providing wide coverage of name and address information; it is used by many other data quality vendors. In June 2009 Address Doctor was acquired by Informatica.	www.addressdoctor.com
Ataccama	Prague-based start-up with a modern data quality suite.	www.ataccama.com
Business Data Quality	UK-based vendor with good government customer references	www.businessdataquality.com
Capscan	London-based provider of address management and data integrity services.	www.capscan.com
Datactics	UK-based vendor specializing in product data quality.	www.datactics.com
Datanomic	Cambridge-based vendor of data quality solutions.	www.datanomic.com
DataFlux	Part of SAS, one of the leading players in data quality.	www.dataflux.com
DataQualityFirst	US start-up whose application lives on top of IBM Quality Stage.	www.dataqualityfirst.com
Datiris	Colorado vendor of data profiling technology	www.datiris.com
Datras	Munich-based vendor with wide ranging data quality functionality	www.datras.de
DQ Global	UK data quality and address verification software	www.dqglobal.com

Vendor	Brief Description	Website
Exeros	California-based vendor specializing in data discovery.	www.exeros.com
Global IDs	New York-based vendor with strong profiling functionality in particular.	www.globalids.com
Help IT Systems	UK vendor of data cleansing technology.	www.helpit.com
Human Inference	Dutch data quality vendor.	www.humaninference.com
IBM	Data quality software from the industry giant.	www.ibm.com
Informatica	California-based vendor, a major player in data quality.	www.informatica.com
Infogix	Illinois-based vendor specializing in controls and compliance.	www.infogix.com
Inquera	Israeli company with innovative approach to product data quality using machine-learning technology based on subject domain experts' knowledge.	www.inquera.com
Innovative Systems	Long-established Pittsburgh-based vendor whose software uses an extensive knowledge base.	www.innovativesystems.com
Intelligent Search	Identity management company now with a more general data quality capability.	www.intelligentsearch.com
Melissa Data	Data quality US vendor with a focus on the Microsoft software environment.	www.melissadata.com
Netrics	New Jersey vendor of impressively accurate matching software	www.netrics.com
Omikron	German data quality vendor with particularly impressive capabilities for structured search and data matching in an international context.	www.omikron.com
Pitney Bowes Business Insight	The data quality vendor formerly known as Group 1, part of the Pitney Bowes group.	www.g1.com
Postcode Anywhere	UK vendor of web-based addressing software.	www.postcodeanywhere.co.uk
QAS Experian	UK-based vendor specializing in customer name and address.	qas.co.uk

Vendor	Brief Description	Website
SAP	The software giant is a major data quality player.	www.sap.com
Satori Software, Inc.	US vendor of address management solutions used by organizations to increase overall address data quality through point-of-entry verification and database cleansing and updating.	www.satorisoftware.com
Silver Creek Systems	Colorado-based vendor of product data mastering software.	www.silvercreeksystems.com
Talend	Paris-based open source data quality software vendor. In September 2009 Talend acquired MDM vendor Amalto Technologies.	www.talend.com
Trillium	Part of Harte Hanks, one of the leading data quality vendors.	www.trilliumsoftware.com
Uniserv	Large German data quality vendor.	www.uniserv.com
X88	Recent UK market entrant specializing in data profiling.	www.x88software.com

Ciant	(www.ciant.com)
Data Lever	(www.datalever.com)
Data Mentors	(www.datamentors.com)
Infosolve	(www.infosolvetech.com)
Intervera	(www.intervera.com)
Irion	(www.iriondq.com)
Ixsight	(www.ixsight.com/)
MSI	(www.msi.com.au)
Stalworth	(www.stalworth.com)
TIQ Solutions	(www.tiq-solutions.com)
Winpure	(www.winpure.com)
Wizsoft	(www.wizsoft.com)

APPENDIX C - THE NORTH WIND TRADER'S DATABASE

NorthWind Traders Tables

Table: Categories

Name	Type	Size
CategoryID	Long Integer	4
CategoryName	Text	15
Description	Memo	-
Picture	OLE Object	-

Table: Customers

Name	Type	Size
CustomerID	Text	5
CompanyName	Text	40
ContactName	Text	30
ContactTitle	Text	30
Address	Text	60
City	Text	15
Region	Text	15
PostalCode	Text	10
Country	Text	15
Phone	Text	24
Fax	Text	24

Table: Employees

Name	Type	Size
EmployeeID	Long Integer	4
LastName	Text	20
FirstName	Text	10
Title	Text	30
TitleOfCourtesy	Text	25
BirthDate	Date/Time	8
HireDate	Date/Time	8
Address	Text	60
City	Text	15
Region	Text	15
PostalCode	Text	10
Country	Text	15
HomePhone	Text	24
Extension	Text	4
Photo	OLE Object	-
Notes	Memo	-
ReportsTo	Long Integer	4

Table: Order Details

Name	Type	Size
OrderID	Long Integer	4
ProductID	Long Integer	4
UnitPrice	Currency	8
Quantity	Integer	2
Discount	Single	4

Table: Orders

Name	**Type**	**Size**
OrderID	Long Integer	4
CustomerID	Text	5
EmployeeID	Long Integer	4
OrderDate	Date/Time	8
RequiredDate	Date/Time	8
ShippedDate	Date/Time	8
ShipVia	Long Integer	4
Freight	Currency	8
ShipName	Text	40
ShipAddress	Text	60
ShipCity	Text	15
ShipRegion	Text	15
ShipPostalCode	Text	10
ShipCountry	Text	15

Table: Products

Name	**Type**	**Size**
ProductID	Long Integer	4
ProductName	Text	40
SupplierID	Long Integer	4
CategoryID	Long Integer	4
QuantityPerUnit	Text	20
UnitPrice	Currency	8
UnitsInStock	Integer	2
UnitsOnOrder	Integer	2
ReorderLevel	Integer	2
Discontinued	Yes/No	1

Table: Ranking_Forms

Name	**Type**	**Size**
Rank_Database	Text	255
Rank_FormName	Text	255
Rank_ColumnName	Text	255
Rank_RowNumber	Single	4

Table: Shippers

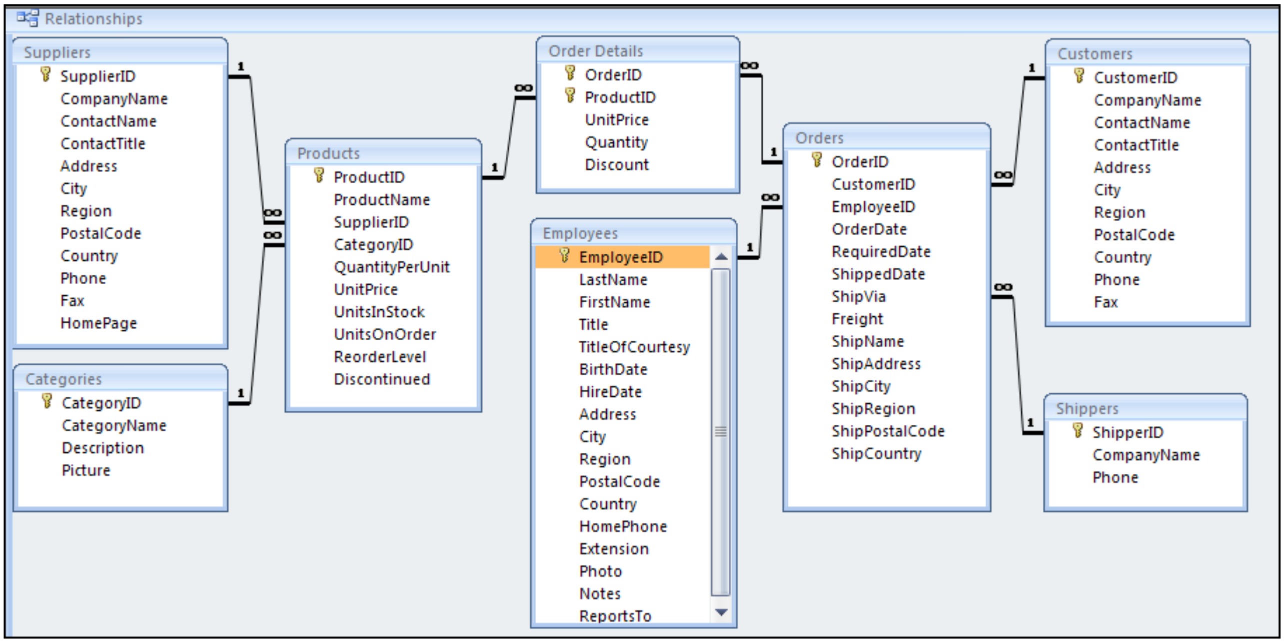
Name	**Type**	**Size**
ShipperID	Long Integer	4
CompanyName	Text	40
Phone	Text	24

Table: Suppliers

Name	Type	Size
SupplierID	Long Integer	4
CompanyName	Text	40
ContactName	Text	30
ContactTitle	Text	30
Address	Text	60
City	Text	15
Region	Text	15
PostalCode	Text	10
Country	Text	15
Phone	Text	24
Fax	Text	24
HomePage	Anchor	-

Appendix D - Relationship Model for North Wind Tables

The North Wind Trader's Database Relationships



Suppliers
SupplierID
CompanyName
ContactName
ContactTitle
Address
City
Region
PostalCode
Country
Phone
Fax
HomePage

Categories
CategoryID
CategoryName
Description
Picture

Products
ProductID
ProductName
SupplierID
CategoryID
QuantityPerUnit
UnitPrice
UnitsInStock
UnitsOnOrder
ReorderLevel
Discontinued

Order Details
OrderID
ProductID
UnitPrice
Quantity
Discount

Employees
EmployeeID
LastName
FirstName
Title
TitleOfCourtesy
BirthDate
HireDate
Address
City
Region
PostalCode
Country
HomePhone
Extension
Photo
Notes
ReportsTo

Orders
OrderID
CustomerID
EmployeeID
OrderDate
RequiredDate
ShippedDate
ShipVia
Freight
ShipName
ShipAddress
ShipCity
ShipRegion
ShipPostalCode
ShipCountry

Customers
CustomerID
CompanyName
ContactName
ContactTitle
Address
City
Region
PostalCode
Country
Phone
Fax

Shippers
ShipperID
CompanyName
Phone

APPENDIX E - DATA NOTES AND AUDIT TRAIL

This facility reflects activities that are collected in every session using an RTF test pad that allows direct export into DOC format.

There is a baseline set of information collected that shows:

```
Date: 9/26/2010 10:28:00 PM
Database: Northwind.mdb, Path: C:\Documents and Settings\XPMUser\Desktop\LOcal Rank
Introduced Error Rate: 10%, General Scalar Value: 1
--------------------------------------------------------------------------------
RANKING PARAMETERS
======================================================
 Query Ranking Multiplier:
--------------------------------------------------------------------------------
 Agreement (single filtered table)                    @ 1.4
 Disagreement (single filtered table)                 @ 1.4
 Agreement and Disagreement (2 tables)                @ 1.3
 Concomitant (up to 4 tables)                         @ 1.2
 Residual   (> 4 tables)                              @ 1.1

--------------------------------------------------------------------------------
Form Query Value:                 1.02
--------------------------------------------------------------------------------
Reports Query Value:              1.03
--------------------------------------------------------------------------------
Reports by Hierarchical Name
--------------------------------------------------------------------------------
Vice President, Sales                  4
Sales Manager                          3
Inside Sales Coordinator               2
Sales Representative                   1

Reference Table for Hierarchical List:    employees, Key Column: reportsto

--------------------------------------------------------------------------------
Business Line Names:
Beverages                   1.205,
Dairy Products              1.192,
Meat/Poultry                1.123,
Confections                 1.131,
Seafood                     1.088,
Condiments                  1.087,
Grains/Cereals              1.085,
Produce                      1.079,
Reference Table for Business Lines:    categories, Key Column: categoryname
----------------------------------------------------------------------------------------

Number of Tables analysed:                       8
Number of Table rows analysed:               3,202
Number of Contestable data elements:        17,397
Number of Queries analysed:                     22
Number of Query rows analysed:              19,336
Number of flagged elements:                  1,654

============================================================================
Product Category Lines with values-at-risk
Category name   Total Sales Value at Risk    Percentage
============================================================================
Beverages       $252,071.15        -$15,797.01    6.27%
Condiments      $99,841.87         -$6,205.18     6.22%
Confections     $156,561.13        -$10,796.02    6.90%
Dairy Products  $223,805.10        -$10,702.15    4.78%
Grains/Cereals  $90,502.89         -$5,241.70     5.79%
Meat/Poultry    $147,531.15        -$15,491.22    10.50%
Produce         $94,067.57         -$5,917.00     6.29%
Seafood         $116,849.95        -$14,411.78    12.33%
----------------------------------------------------------------------------------------
```

In addition, other results are noted depending upon the functions selected.
Each session offers the option to save the notes as a discrete named file that can be imported into Word

APPENDIX F - ATTACHED CD

These appendices are also listed on the attached CD. If the CD is missing then the contents can be made available by emailing a request to: CharlesPalmer@charlespalmer.net.

The contents of the CD are:

1. An EXCEL Spreadsheet showing the initial ranked file for all contestable data elements. The spreadsheet has a series of TABs showing the differences between the contestable data elements seeded with 1% through to 10% in increments of 1%.

2. The original North Wind Trader's Database as released by Microsoft

3. An MS WORD document describing the software artifact

4. An MS WORD document describing the Literature Research Tool, LiMITS

5. A listing of all literature reviewed