



# Ministry of Forests and Range Data Quality Framework

## A Guide for Business Data Quality

This framework provides an approach for MoFR to analyse and report on the quality of Ministry data, and provides methodology to improve data quality.

It should be used as a key reference for future phases or any interim steps that may proceed.

Framework Completion Date – March 2006



Tom Fulton and Jeremy Janzen  
Data Administration  
Ministry of Forests and Range



The Center for Data Quality, Inc  
[www.c4dq.com](http://www.c4dq.com)  
Richard W. (Dick) Paar  
Executive Vice President



Dr Marla Weston  
Data Quality Specialist  
Weston Spatial Business Solutions

## DOCUMENT HISTORY

<i>Document Version Number</i>	<i>Date Issued</i>	<i>Comments</i>	<i>Issued By</i>
0.01	February 14, 2006	C4DQ prepared Table of Contents for MoFR approval	Center for Data Quality
0.02	February 21, 2006	First cut at some of the easier framework sections for MOFR review and comment	Center for Data Quality
0.03	March 5, 2006	Reformatting of document. Update of section 2.2. Rewrite of definitions and inclusion of spatial component.	Weston Spatial Business Solutions
0.04			
0.05	March 13, 2006	Updated sections 2 and 3	MoFR – Tom Fulton
0.06	March 28, 2006	Reformatted report, re-ordered sections.	Weston Spatial Business Solutions
Final Draft	March 31, 2006		Weston Spatial Business Solutions
Final		Final review and updates.	MoFR – Tom Fulton and Jeremy Janzen

## Acknowledgements

We would like to thank staff from a number of major MoFR business areas who through interviews provided excellent insight into data quality concerns and initiatives within their areas of expertise:

- Dona Stapley, Data Integrity Analyst HTH with Bruce Bell, IMG, MoFR. Dona is project lead on a multi-year, multi-million dollar cleanup project for Forest Tenure data. She provided knowledge of the forest tenure business, problems that caused the need for data quality cleanup, approaches to resolve the DQ problems and to keep from repeating them.
- Gail Brewer, Manager of Program Planning & Business Practices, BCTS. Gail explained the BC Timber Sales business and the issues and impact data quality can have on that business. BCTS has the desire to achieve ISO Quality Management certification which Gail has investigated and will pursue.
- Jon Vivian, Manager of Inventory Program, FAIB, MoFR along with Tim Salkeld and John Wakelin. Jon, Tim and John explained the inventory program, data quality improvement initiatives with associated costs and progress made through electronic submission of vegetation inventory updates. While the Vegetation Inventory data is the best available, they would like to improve it to significantly increase the usefulness but costs are prohibitive.
- Ralph Winter, Silviculture & Management Administrator FPB along with consultants Scott Killam and Tony Dellavilova. Ralph, Scott and Tony explained reforestation silviculture obligations, data quality improvement initiatives undertaken for silviculture obligation data, progress made through electronic submission, and the importance of data quality in tracking these legal obligations within RESULTS.
- Thomas Chen, Quality Control and Data Management Specialist (FREP) FPB. Thomas' role is quality and he has assisted all along with the data quality project contributing research and expertise.

## Executive Summary

The BC Ministry of Forests and Range (MoFR) is committed *To protect, manage and conserve forest and range values through a high performing organization*. A key requirement to achieving this mission is the capture and management of high-quality information necessary to support the diverse activities of MoFR. Ministry business areas have undertaken various initiatives over the last few years driven by the need for quality data. For example, RTE has undertaken the forest tenure data cleanup, BCTS has started work towards ISO 9000 Quality Management certification, FAIB has started work towards standardized Data Management Plans for all data, and FPB has undertaken Results data cleanup.

This Data Quality Framework is intended to support these initiatives and others by taking a broad Ministry approach to guide data quality analysis, reporting and improvement. Meeting business and legal requirements, and protecting the significant investment made in MoFR data quality is a concern of Ministry executives, Data Custodians, management and staff.

The ultimate objective of this framework is to provide a roadmap to guide the Ministry in moving forward with data quality analysis, reporting and improvement.

MoFR is known for its leadership. The Data Quality Improvement initiative is another opportunity to lead government and industry in an emerging business discipline.

### ***Recommended Direction***

#### **SHORT-TERM**

In order to provide an enduring roadmap for data quality, it is essential to learn more about the current state of data quality from a broad perspective within the Ministry. The quickest and most cost effective method to gain a better understanding in the short-term is to implement a data quality analysis and reporting tool and to start with some small initial projects. Data quality analysis and reporting tools allow automatic validation and monitoring of the quality of a wide range of enterprise data. Such a toolset can be used to assess data quality within any interested business area. It can also be used to support quality management certification such as the ISO quality program. The data quality report cards produced by these tools will allow business areas to identify problem areas within their data. With this information at hand, the business areas may then determine which problems, if any, require attention. It will also allow business areas to learn more about their data, with an ultimate objective to promote a data quality culture within the Ministry.

## **LONG-TERM**

The long-term objective of this framework is to support a Ministry wide data quality improvement program which will achieve the right level of data quality to meet Ministry, forest industry, BC government and public business needs at a reasonable cost. This long-term objective may be best stated as a vision statement.

### ***Vision Statement for Data Quality at MoFR***

*The Ministry of Forests and Range records the quality of data in every business area, proudly shows where high data quality is meeting business needs, and continuously demonstrates improvement in the quality of data where required.*

# Data Quality Framework Table of Contents

<b>EXECUTIVE SUMMARY .....</b>	<b>I</b>
RECOMMENDED DIRECTION .....	I
<i>Short-Term</i> .....	<i>i</i>
<i>Long-Term</i> .....	<i>ii</i>
VISION STATEMENT FOR DATA QUALITY AT MOFR .....	II
<b>1. INTRODUCTION .....</b>	<b>5</b>
1.1. FRAMEWORK OBJECTIVES .....	5
1.2. WHY ANALYSE OR IMPROVE DATA QUALITY? THE DRIVING FORCES .....	6
1.3. STAKEHOLDERS .....	7
1.4. SCOPE OF FRAMEWORK .....	7
<b>2. KEY CONCEPTS.....</b>	<b>8</b>
2.1. WHAT IS DATA QUALITY?.....	8
2.2. WHAT IS DATA QUALITY TO MOFR? .....	9
<b>3. RECOMMENDED DIRECTION .....</b>	<b>14</b>
3.1. WHERE TO BEGIN .....	<b>ERROR! BOOKMARK NOT DEFINED.</b>
3.2. SHORT-TERM .....	14
3.3. NEAR-TERM .....	15
3.4. LONG-TERM .....	15
<b>4. PROPOSED METHODOLOGY .....</b>	<b>16</b>
4.1. IDENTIFY PROBLEMS .....	17
4.2. ANALYZE SOLUTIONS .....	18
4.3. IMPLEMENT CHANGES .....	19
4.4. MONITOR RESULTS .....	20
4.5. RELATIONSHIP TO POPULAR QUALITY STANDARDS .....	20
4.6. WHAT TO MEASURE? .....	21
<b>5. THE MOFR DATA QUALITY MANAGEMENT ROLES .....</b>	<b>23</b>
5.1. CURRENT MANAGEMENT PROCESSES AT MOFR .....	23
5.2. OVERALL MINISTRY APPROACH .....	25
5.3. LONG-TERM MANAGEMENT GOALS.....	26
<b>6. DATA QUALITY MANAGEMENT TOOLS .....</b>	<b>27</b>
6.1. EXISTING TOOLS .....	27
6.2. REQUIRED TOOLS / INDUSTRY STANDARD TOOLS .....	27
6.3. ATTAINING SUSTAINABLE DATA QUALITY LEVELS .....	31
6.3.1. IMPLEMENTING DATA QUALITY CERTIFICATION.....	32
6.3.2. INTERNATIONAL STANDARDS.....	32
6.4. ORGANIZATIONAL IMPACT .....	33
6.5. BEYOND DATA QUALITY MANAGEMENT TOOLS .....	34
<b>7. APPENDIX A – REFERENCES.....</b>	<b>35</b>
<b>8. APPENDIX B – ALPHABETICAL LISTING OF ATTRIBUTE DATA QUALITY DIMENSIONS .....</b>	<b>36</b>
8.1. A COMPARISON OF ATTRIBUTE DIMENSIONS .....	44
8.2. MOFR ATTRIBUTE DATA QUALITY DIMENSIONS DETAILS .....	45
8.2.1. ACCURACY.....	45
8.2.2. PRECISION .....	45

8.2.3.	CONSISTENCY.....	46
8.2.4.	BELIEVABILITY / RELIABILITY .....	47
8.2.5.	COMPLETENESS.....	47
8.2.6.	CURRENCY / TIMELINESS .....	48
8.2.7.	CONSISTENT REPRESENTATION.....	48
8.2.8.	APPROPRIATENESS .....	49
8.2.9.	ACCESSIBILITY.....	49
8.2.10.	AVAILABILITY.....	49
8.2.11.	METADATA .....	50
<b>9.</b>	<b>APPENDIX C – ALPHABETICAL LISTING OF SPATIAL DATA QUALITY DIMENSIONS</b>	
	<b>51</b>	
9.1.	A COMPARISON OF SPATIAL DIMENSIONS .....	55
9.2.	MOFR SPATIAL DATA QUALITY DIMENSIONS DETAILS.....	56
9.2.1.	POSITIONAL ACCURACY .....	56
9.2.2.	THEMATIC ACCURACY .....	56
9.2.3.	COMPLETENESS.....	57
9.2.4.	LOGICAL CONSISTENCY .....	57
9.2.5.	TEMPORAL ACCURACY .....	57
<b>10.</b>	<b>APPENDIX D – DETAILS OF BUSINESS RULES .....</b>	<b>59</b>
<b>11.</b>	<b>APPENDIX E – DATA QUALITY METHODOLOGIES.....</b>	<b>64</b>
<b>12.</b>	<b>APPENDIX F – INTERVIEWS .....</b>	<b>ERROR! BOOKMARK NOT DEFINED.</b>

## Table of Figures

<i>Figure 1. Data Quality Business Rules – Measuring Data Quality.....</i>	<i>14</i>
<i>Figure 2. Data Quality Improvement Cycle .....</i>	<i>17</i>
<i>Figure 3. Identify Problems.....</i>	<i>17</i>
<i>Figure 4. Analyze Solutions.....</i>	<i>18</i>
<i>Figure 5. Implement Changes .....</i>	<i>19</i>
<i>Figure 6. Monitor Results.....</i>	<i>20</i>
<i>Figure 7. Data Quality Assessment.....</i>	<i>28</i>

## Table of Tables

<i>Table 1. Attribute Data Quality Dimensions .....</i>	<i>11</i>
<i>Table 2. Spatial Data Quality Components.....</i>	<i>12</i>
<i>Table 3. A high-level summary of some representative information/data quality management methodologies.....</i>	<i>21</i>
<i>Table 4. Attribute dimension automations.....</i>	<i>22</i>
<i>Table 5. Automated Data Quality Tool Requirements.....</i>	<i>29</i>

# 1. Introduction

Data quality can often seem a hazy concept, but the lack of data quality severely hampers the ability of organizations to effectively accumulate, manage, and make proper use of enterprise-wide knowledge. Quality is an issue for any kind of information.

The BC Ministry of Forests and Range is committed *To protect, manage and conserve forest and range values through a high performing organization*. A key requirement to achieving this mission is the capture, management and use of the high-quality information necessary to support the diverse activities of MoFR. Since it is impossible derive high-quality information from data of suspect or low quality, it is important to adopt an organizational culture in which all employees, data suppliers and other stakeholders make improving the quality of data a priority. This data quality framework is being developed as a future direction for all Ministry of Forests and Range operations. This framework is proposing a data quality improvement methodology that is:

- enterprise-wide in scope,
- systematic in approach,
- measurable in terms of data quality dimensions, and
- understandable and repeatable.

## 1.1. Framework Objectives

### WHY THIS FRAMEWORK?

Ministry business areas have undertaken various initiatives over the last few years driven by the need for or associated with data quality. RTE has undertaken the forest tenure data cleanup, BCTS has started work towards ISO 9000 Quality Management certification, FAIB has started work towards standardized Data Management Plans for all data, and FPB has undertaken Results data cleanup.

The Data Quality Framework is intended to support these initiatives and others by taking a broad Ministry approach to guide data quality analysis, reporting and improvement.

### WHAT ARE THE OBJECTIVES?

The objective of the framework is to provide the roadmap to guide the Ministry in moving forward for data quality analysis, reporting and improvement methodology. It is also to support a cost justified data quality improvement program which will achieve the right level of data quality to meet Ministry, forest industry, BC government and public business needs.

The ultimate objective is to encourage a constant striving for even higher quality data within the Ministry as well as by the Ministry's various partners and stakeholders. In other words, instil a culture of data quality.

## 1.2. Why Analyse or Improve Data Quality? The Driving Forces

Missing, incomplete or inaccurate data has caused MoFR serious business problems, and the Ministry has invested significant time and money in improving data quality. Some examples are<sup>1</sup>:

- BCTS has started in the last couple of years to clean up the timber sales data held in GENUS.
  - Staff believe that improved data quality would result in increased revenue from some of their approximately 20 percent of the MoFR timber sales.
  - Data quality is also important because of downstream use. The data is fed into: the Revenue Branch costing and pricing systems, FTA, RESULTS, the Roads Management system, and the Land and Resource Data Warehouse as part of Land Information BC.
  - BCTS also has the desire to achieve ISO Quality Management certification. Data quality improvement is a significant component of the ISO Quality Management program.
- From 1996 to 2002, \$700,000 was invested by MoFR in cleaning up Sample Plot data to support greater accuracy in estimation of the provincial Vegetation Inventory.
- Recent legislative changes have resulted in all legal silviculture obligations being tracked through the RESULTS system. As the only record of these obligations, quality of this data is essential for legal requirements. FPB plans to invest \$450,000 in data cleanup over the next few years.
- Business problems were experienced due to forest tenure data quality, and HTH and the BC Government have invested approximately \$5,000,000 in cleaning up forest tenure data over the last 3 years due to missing and/or inaccurate data.

Meeting business and legal requirements, and protecting the significant investment made in MoFR data quality is a concern of Ministry executives, Data Custodians, management and staff.

### Why would we do this?

- To support a desire by Data Custodians to understand the quality of their data so they can demonstrate high quality where appropriate and plan for improvement where required.
- To support a desire by Data Custodians to have an on-going understanding of the quality trends within their data so significant investments made in data quality improvements to date are not jeopardized.
- To support the desire by some Data Custodians to seek ISO Quality Management certification by providing data quality analysis and reporting capabilities and improvement methodology required as part of the ISO Quality Management program.
- To meet data quality requirements necessary to meet legal obligations.

---

<sup>1</sup> Data Quality interviews were held with BCTS, FAIB, FPB and HTH staff.

MoFR is not alone in the movement towards improved data quality. According to Forrester Research<sup>2</sup>, the Information Quality market is growing at 14% annual growth rate and the Information Quality Software market is growing at an annual rate of 17%. Other standard reasons for improving data quality in MoFR and organizations of every type across Canada and around the world are:

Operations Cost-Efficiency	Running operations built upon suspect or bad data has a negative impact on cost-efficiency. Also the time and money spent cleaning poor quality data is time and money taken away from other more productive activities.
Customer Service	If data supporting customer activities is missing, inaccurate or unavailable, the ability to service and satisfy customer requirements is greatly compromised.
Error Detection And Correction	Unless an organization has a formal data quality initiative to systematically detect and correct problems, error detection tends to be hit or miss, accidental, or through finite data cleanup projects. Often errors become special cases requiring manual correction. If the source of the error is not determined and improved, then the error is destined to recur. The costs of this type of adhoc or short term error correction can be significant
Ability To Make Valid Decisions	The quality of the data becomes a significant issue when Ministry management is unaware of data quality problems that may affect the accuracy of the vital business decisions
Loss Of Employee Confidence	Poor data quality can be a source of extreme frustration and dissatisfaction within the workforce causing “self-defense” spreadsheets or databases to be created by employees to track information needed to perform job duties. Duplicate databases and data spreadsheets only add to workload, as well as have negative impacts on data integrity and can lead to inconsistency of business results.
Data Profile Raised Through Data Warehousing	Data warehouses, accumulating decades of operational data, uncover data problems with significant regularity. A leading cause of failure of an implemented warehouse to meet the expectations of the planners is incomplete or inaccurate data.

### 1.3. Stakeholders

- Information Governance Council
- Data Custodians
- Data Standards Managers
- Business Area Experts dealing with business data and the quality associated with those data
- Sponsor: Chief Information Officer

### 1.4. Scope of Framework

The purpose of this framework is to provide overall Ministry guidance in terms of data quality analysis and improvement. It is intended to be the first step for moving forward towards putting Ministry wide analysis and improvement infrastructure in place.

<sup>2</sup> <http://www.forrester.com/my/1,,1-0,FF.html>

The framework will be a dynamic document in terms of what is learned from the results of Ministry data quality improvement projects. Once analysis tools are put in place within the next phase of the project, and business areas embark on data quality analysis and improvement projects, the framework will be updated to reflect analysis and methodology improvements.

## 2. Key Concepts

### 2.1. What is Data Quality?

Quality expert Larry English provides the following definition for data quality:

What, then, is quality? Total Quality Management provides a useful definition of quality: “consistently meeting customer’s expectations.”<sup>3</sup>

This definition is echoed in most other definitions of data quality. For example, the United States Census Bureau defines data quality as “... fitness for use. We are guided by the needs of our customers to ensure our data products are fit for their use.”<sup>4</sup> Statistics Canada also defines data quality in terms of “fitness for use”. They further state that “Whether data and statistical information are fit for use depends on the intended uses and on intrinsic characteristics of the data or information.”<sup>5</sup>

But is data quality the same as information quality? According to quality expert Larry English “information is data in context.” In other words, it is usable data. Larry English defines information quality as follows:

There are two significant definitions of information quality. One is its inherent quality, and the other is its pragmatic quality. Inherent information quality is the correctness or accuracy of data. Pragmatic information quality is the value that accurate data has in supporting the work of the enterprise. Data that does not help enable the enterprise accomplish its mission has no quality, no matter how accurate it is.<sup>6</sup>

Information or data quality, as defined by Larry English, is not an esoteric notion. It directly affects the effectiveness and efficiency of business processes. For those who use the data or information in any form as part of their job function or in the course of performing a process, quality data is essential. These “customers of information” or “knowledge workers” are best able to discern whether data has quality or not based upon how well the data supports their ability to do their jobs. Virtually all employees are knowledge workers. From the executives who make the decisions to the order entry clerks who create orders. All these knowledge workers need quality data. Their needs may be summarized as<sup>7</sup>:

---

<sup>3</sup> English, L., 1999. *Improving Data Warehouse and Business Information Quality: Methods for reducing costs and increasing profits*. John Wiley & Sons, Inc. p. 17.

<sup>4</sup> Census Bureau Principle Version 1.2. Jan 5, 2006. Definition of Data Quality. ([http://www.census.gov/quality/P01-0\\_v1.2%20Definition%20of%20Quality.htm](http://www.census.gov/quality/P01-0_v1.2%20Definition%20of%20Quality.htm))

<sup>5</sup> Statistics Canada, March 31, 2000. Policy on Informing Users of Data Quality and Methodology. (<http://www.statcan.ca/english/about/policy/infousers.htm>)

<sup>6</sup> English, L., 1999. *Improving Data Warehouse and Business Information Quality: Methods for reducing costs and increasing profits*. John Wiley & Sons, Inc. p. 17. p. 32.

<sup>7</sup> *Ibid*, p. 31.

<b>Quality Characteristic</b>	<b>Knowledge Worker Benefit</b>
The <i>right data</i>	The data I <i>need</i>
With the right <i>completeness</i>	<i>All</i> the data I need
In the right <i>context</i>	Whose <i>meaning</i> I know
With the right <i>accuracy</i>	I can <i>trust</i> and rely on it
In the right <i>format</i>	I can <i>use</i> it <i>easily</i>
At the right <i>time</i>	<i>When</i> I need it
At the right <i>place</i>	<i>Where</i> I need it
For the right <i>purpose</i>	<i>I can accomplish our objectives and delight our customers.</i>

The measurement of these quality characteristics is more problematic. The issue is summed up by MIT’s Total Data Quality Management research program,

Our studies have revealed that data quality has a number of dimensions for data users, including accuracy, believability, relevancy, and timeliness. A clear and uniform articulation of data quality metrics is needed. In fact, even a relatively obvious dimension, such as accuracy, does not have a sufficiently robust definition to make techniques apparent as to how to measure the accuracy of data.<sup>8</sup>

Because good data quality is frequently defined in terms of “fitness for use”, it can be difficult to delineate fitness when there are no metrics against which to measure it. The assessment of any data’s levels of quality can be done in the context of what are referred to as the dimensions of data quality. It is through the process of classifying requirements and setting measurement goals that data quality can be improved.

## **2.2. What is Data Quality to MoFR?**

Data quality as “fitness for use” is essential to MoFR. As noted above, it is not always easy to measure what this means. Data quality dimensions can be used to identify business area data quality requirements, to measure levels of data quality and to identify the gaps and opportunities for data quality improvement.

Appendices B and C of this framework lists forty-four possible candidate attribute dimensions and thirty-two candidate spatial dimensions for evaluating the quality of data. Through interviews and discussions with various Ministry personnel, a subset of the data quality dimensions that are most relevant to MoFR has been identified.

### **WHAT IS MEASURED?**

#### **ATTRIBUTE DIMENSIONS**

A total of eleven (11) attribute dimensions were selected out of the complete list of 44 possible data quality dimensions for attribute data. The importance of each of the dimensions will vary with

---

<sup>8</sup> <http://web.mit.edu/tdqm/www/about.shtml>

business area. The selected dimensions of data quality can be grouped into four main categories as defined by Lee et al.<sup>9</sup>:

1. **Intrinsic data quality** – This category implies information (data) has quality in its own right.
2. **Contextual data quality** – This category highlights the requirements that information quality must be considered within the context of the task at hand; it must be relevant, timely, complete and appropriate in term of amount, so as to add value.
3. **Representational data quality** – For this category, the system must present information in such a way that it is interpretable, easy to understand, easy to manipulate, and is represented concisely and consistently
4. **Accessibility data quality** – This category states the system must be accessible but secure.

A comparison of all 44 possible dimensions based on the first four categories is provided in Appendix B (8.1). The eleven dimensions identified as most generally relevant to MoFR are listed below. Again, it is critical to note that the relative importance of each of the listed dimensions will vary with individual business areas.

Category	Dimension	Definition
<b>Intrinsic</b>	Accuracy	Accuracy refers to how closely the data value agrees with the correct or “true” value.
	Precision	Precision is the ability of a measurement or analytical results to be consistently reproduced, or the number of significant digits to which a value has been measured or calculated. One can simultaneously be extremely precise and totally inaccurate.
	Consistency	Data consistency refers to the common definition, understanding, interpretation and calculation of a data element.
	Believability / Reliability	The degree of credibility or trustworthiness of the information.
<b>Contextual</b>	Completeness	Completeness refers to the expectation that certain attributes are expected to have assigned values in a data set. Also includes missing associated records.
	Currency / Timeliness	Currency refers to the degree to which information is current with the world that it models.
<b>Representational</b>	Consistent representation	This dimension refers to whether data elements or symbols are consistent with a defined standard or style.
	Appropriateness	Appropriateness is the dimension that categorizes how well the format and presentation of the data matches the users' needs.
<b>Accessibility</b>	Accessibility	Accessibility refers to the degree of ease of access to information, as well as the breadth of access (whether all the information can be accessed).
	Availability	Data is available when it is needed.
	Metadata	Presence of an enterprise-wide metadata

<sup>9</sup> Lee, Y.W., D.M. Strong, B.K. Kahn, & R.Y. Wang, 2002. AIMQ: a methodology for information quality assessment. *Information & Management* 40: 133-146.

Category	Dimension	Definition
		framework and support policies.

**Table 1. Attribute Data Quality Dimensions**

A more detailed examination of the measurement of these dimensions along with examples is provided in Appendix B.

### SPATIAL DIMENSIONS

The dimensions examined above focus on defining the quality of attribute data, but a large portion of the data managed by MoFR is spatial data. Attribute data defines what is present, while spatial data defines where it is located. It is important to note that spatial data is almost always accompanied by attribute data. As a result, spatial data sets and their attributes are subject to the same data quality dimensions given in Table 1 above. However there are unique aspects to spatial data which require special data quality dimensions.

Unlike attribute data, spatial data must of necessity be approximate, since it is impossible to capture all of the infinite complexity of the Earth’s surface in any form, whether as a map, a digital database, or a narrative. Many efforts have been made over the past decades to define spatial data quality, and to identify its core elements<sup>10</sup>. The original definition of spatial data quality by the Spatial Data Transfer Standard (Digital Cartography Data Standards Task Force, 1988<sup>11</sup>) had a huge influence on subsequent standards works including that of the ISO standards Technical Committee 211. The spatial data quality components or dimensions defined in the ISO/TC 211 (19115) standard that are most applicable to MoFR are:

Spatial Dimension <sup>12</sup>	Definition
Positional Accuracy	The closeness of locational information (usually coordinates) to the true position on the earth.  <i>For example:</i> Administrative boundaries are determined to be inaccurate when compared with survey results.
Thematic Accuracy	The accuracy of quantitative attributes (e.g. 30% Douglas Fir) and the correctness of non-quantitative attributes (e.g. wrong BEC Zone for Douglas fir) and the classification of features and their relationships.  <i>For example:</i> For timber sales, conducting a thematic accuracy assessment involves collecting reference (ground truth) information and then comparing these data to the data collected by the Timber Cruise.  A thematic accuracy assessment has three major components: <ul style="list-style-type: none"> <li>o Determining the sample unit. What will the field observation consist of? Often this is a physical area of a certain size or a certain number of pixels for remote sensing.</li> <li>o Selecting a sampling approach. Often this is some variation of a random</li> </ul>

<sup>10</sup> Shi, W., P.F. Fisher & M.F. Goodchild (eds), 2002. *Spatial Data Quality*. (London: Taylor & Francis).

<sup>11</sup> Digital Cartographic Data Standards Task Force, 1998. The proposed standard for digital cartographic data. *The American Cartographer*, 15(1) pp. 21-137.

<sup>12</sup> Note: In the ISO/TC 211 standard, “dimensions” are referred to as “components”.

<b>Spatial Dimension<sup>12</sup></b>	<b>Definition</b>
	<p>sample selection process and is stratified by habitat or species.</p> <ul style="list-style-type: none"> <li>○ Reporting the results. This involves creating an error matrix and reporting overall accuracy, as well as individual habitat or species accuracies.</li> </ul>
Completeness	<p>Are all required objects included within the database? Completeness in the spatial sense can be measured in space, time, or theme.</p> <p><i>For example:</i> For example, looking at all the cut blocks in British Columbia. <b>Spatial completeness:</b> The database only contains cut blocks on Vancouver Island (not entire province). <b>Temporal completeness:</b> The database only contains cut blocks since 2002. <b>Thematic completeness:</b> The database only contains cut blocks on Crown Land.</p>
Logical Consistency	<p>The degree of conformance of a geographical data set with respect to the internal structure given in its specification.</p> <p><i>For example:</i> Can't have a forest tenure with cutting permit over a lake.</p>
Temporal Accuracy	<p>The accuracy of the temporal attributes and temporal relationships of features. Includes the correctness of the temporal references of an item (reporting of error in time measurement), the correctness of ordered events or sequences if reported, and the validity of the data with respect to time.</p> <p><i>For example:</i> Temporal coordinates are often only implicit in geographical data, e.g., a time stamp indicating that the entity was valid at some time. Often this is applied to the entire database (e.g., a map dated "1995"). More realistically, temporal coordinates are the temporal limits within which the entity is valid (e.g., Pothole Q54D-35-021 existed between 2/12/96 and 8/9/96). Temporal accuracy is not the same as "database time", which is the time the information was entered into the database. Temporal accuracy is not the same as "currentness" (or up-to-dateness) which is actually an assessment of how well the database specification meets the needs of a particular application. A database can be temporal accurate but still out of date; historical applications depend on such data.</p>

**Table 2. Spatial Data Quality Components**

These spatial data quality components along with methods for measuring them are examined in more detail in Appendix C.

#### HOW ARE DIMENSIONS MEASURED? THE BUSINESS RULES

Many of the attribute and spatial data quality dimensions require a contextual knowledge to ascertain quality. However, through the definition of data quality business rules, it is possible to derive a means to both assess the current state and measure the progress in improvement of the data's quality.

Data quality business rules provide a formal way to describe what is expected in terms of data quality. These rules might be as simple as the definition of a primary key being unique and not null, or as complex as saying that if harvesting is complete for a cut block, then an opening must exist except under special conditions. The advantage of defined business rules is that they can lead to an automated discovery of data quality.

## AUTOMATED VS MANUAL DATA QUALITY ANALYSIS

Consolidating business and data quality rules is a way to capture and control strategic knowledge. For an organization such as MoFR there is a significant advantage to this consolidation using an automated rules system to implement those rules. Advantages of using an automated rule-based approach include:

- Efficiency in policy automation
- Opportunities for reuse
- More consistent application of business rules
- Ensure consistent data quality analysis capabilities across MoFR
- Greater data quality buy-in from personnel if manual processes are minimized.

Defined business rules by individual business areas will play a central role to an automated process. A simplified view of this process might be:

- The individual business areas determine the relevant business rules and target data.
- The business rules are captured in a corporate data quality analysis and reporting tool.
- The captured rules are applied against the target data to produce a report card of data quality. In many cases, this will be an executive level summary.
- Once the report card is available, business experts drill down to find the details of the data quality problems.
- The experts then either correct the problems and re-check the data, or determine that there are no problems or that any problems are within an acceptable tolerance level. The data is then passed to a final quality set.

A generalized overview of the automated process is given in Figure 1.

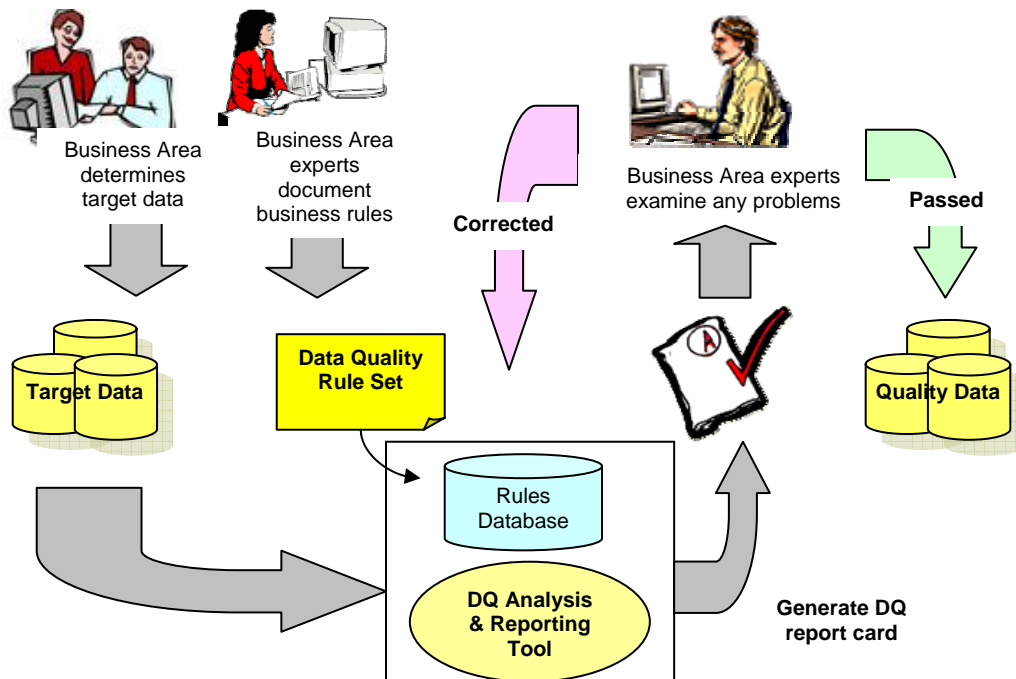


Figure 1. Data Quality Business Rules – Measuring Data Quality

An examination of some representative business rules is provided in Appendix D.

Not all data quality dimensions can be evaluated using a software tool. For example, the intrinsic dimension of believability or reliability. The degree of credibility or trustworthiness of the information may be a judgment call. Often, the first indication that there are issues with the reliability of a data set comes in the form of feedback (often negative) from knowledge workers or customers of the data set.

Similar issues exist with the dimensions of accessibility and availability. Is a data set easy to access and available when it is needed? Users of that data set know very well when it is not, but this is difficult to measure with an automated tool.

The measurement of dimensions which require user feedback or a form of “ground truthing” to ensure accuracy is much more complex. While these dimensions are important, they can be time consuming and expensive to assess for data quality.

Using a data quality tool with the dimensions that can be measured in an automated way will, in the first instance, provide a better return on investment. In addition, the ability to identify data quality issues more quickly and easily will likely lead to greater staff buy-in of a data quality management approach. Also by correcting the data quality problems highlighted by an automated tool, issues around reliability or believability among others may also be addressed.

### **3. Recommended Direction**

The purpose of this framework is to provide guidance for improvement of MoFR data quality. This begins with the ability to analyze data quality. The knowledge gained from the analysis can then be used to improve data quality where required.

#### ***3.1. Short-Term Improvement***

The critical first-step is to obtain a better understanding of the current state of data quality within Ministry of Forests and Range. To start this, the implementation of a data quality analysis and reporting tool is vital. The data quality analysis and reporting tools are designed and built to provide an effective means to test business rules both in the initial analysis phase and as an ongoing monitor of data quality. The generated reports and associated “report cards” will allow business areas to identify problems within their data and then determine which problems, if any, require attention. It will also allow business areas to learn more about their data, with an ultimate objective to promote a culture of data quality within the Ministry.

The steps for the short-term are:

1. Implement a data quality analysis and reporting tool.
2. Make the data quality analysis and reporting tool available to interested business areas.

3. Have the business areas select a data set(s) for a pilot project and then document some basic business rules for use with the data quality analysis and reporting tool.
4. Using the tool, identify and document problem areas, if any, with the selected data set(s).
5. Prioritize needs and levels of effort with respect to correcting any identified data quality problems. The initial focus should be on those problems which are relatively simple to fix but provide maximum data quality improvements.
6. Publicize results to other business areas, update the framework and celebrate successes.
7. Build to establishing best practices and work to entrench the culture of quality.

A detailed examination of the recommended methodology to achieve the short-term (as well as near- and long-term) goals is provided in section 4.

### ***3.2. Near-Term Improvement***

Once the short-term pilot projects are complete, more business areas should be encouraged to participate. The lessons learned from the pilot projects should be documented and made available to all other business areas. It will be important to ensure the data quality framework, and related best practices and standards are updated as required.

The goal in the near-term will be to improve the quality management system within the Ministry and to look for other opportunities for improvement.

### ***3.3. Long-Term Improvement***

The long-term objective of the data quality framework is to support a data quality improvement program which will achieve the right level of data quality to meet Ministry, forest industry, BC government and public business needs at a reasonable cost.. This program should be implemented in all aspects of the data life-cycle including:

- The Information Governance Council to act as a forum to:
  - Instil, promote and foster a cultural change that emphasizes that good quality data is an all-pervasive mindset in the organization
  - Act as the champion for ensuring the necessary funding to ensure success of the data quality improvement program
  - Discuss and coordinate data quality improvement activities in all divisions
  - Be the final arbiter of quality-related conflicts across divisions
  - Be the source for and coordinator of required data quality training for all personnel throughout the organization
  - Empower employees at all levels of the organization to improve the quality of data under their purview
- Development of a long-term plan to systematically assess the quality of the data in all key Ministry business areas and determine corrective action requirements that not only clean up current issues but also correct the upstream policies, processes and procedures that allow data quality issues to occur
- Implementation of ongoing automatic monitoring and trending of the quality of key business data to ensure that the data remains of high quality over time.

- Entry into agreements with data suppliers regarding the minimum acceptable levels of measured quality of those data feeds such that data below standard will be rejected.
- Implementation of automated data quality monitors on all data warehouse feeds to ensure that only high quality data is being loaded
- Ensure that the System Development Life Cycle properly uses best practices to achieve data quality in new development
- Full support for existing data standards, and creation and implementation of new standards such that each data element is defined and used in the same way throughout the organization.

The ultimate aim is the realization of the vision statement for data quality within the Ministry. This vision statement is:

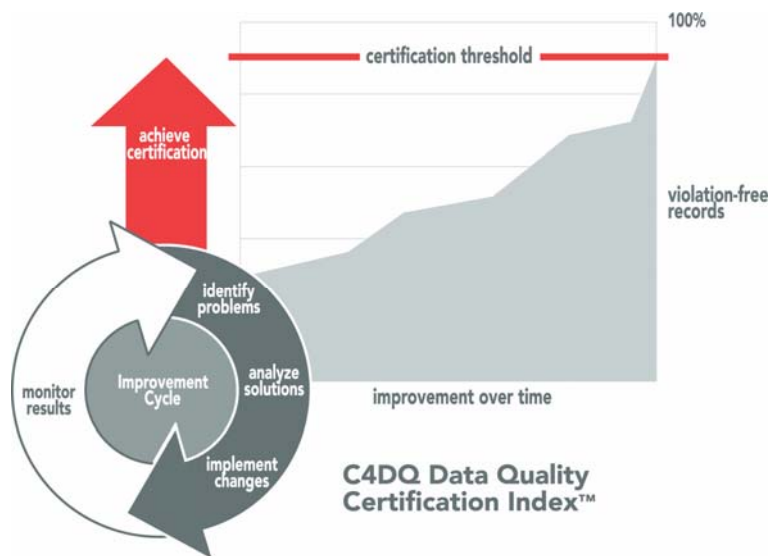
***Vision Statement for Data Quality at MoFR***

*The Ministry of Forests and Range records the quality of data in every business area, proudly shows where high data quality is meeting business needs, and continuously demonstrates improvement in the quality of data where required.*

**4. Proposed Methodology**

Many of the fundamental principles of data quality are not new and have been applied in numerous areas including manufacturing and service quality. There are many quality approaches, but the concept of continuous quality improvement first proposed by Walter Shewhart of Bell Laboratories and later espoused by W. Edwards Deming is the basis of most of the current data quality improvement methodologies.

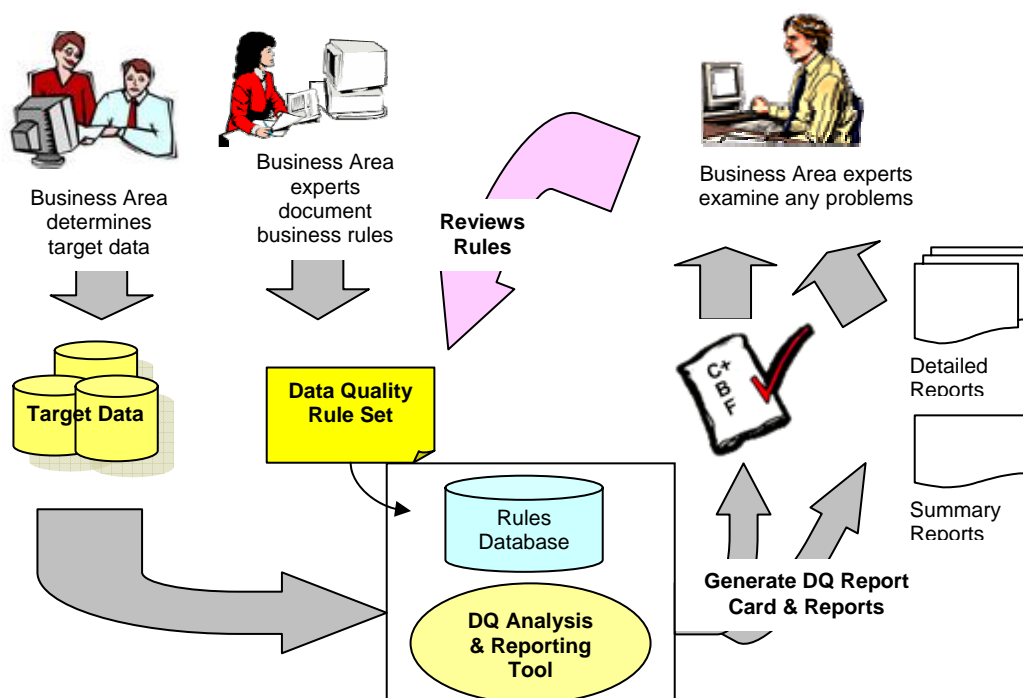
The methodology proposed for MoFR is follows this concept of continuous quality improvement and is outlined in Figure 2. The cycle begins with problem identification, moves through analyzing the solutions and implementing the changes, and finally to monitoring the results on a continuous basis.



**Figure 2. Data Quality Improvement Cycle**

In the short-term, it is proposed that a data quality analysis and reporting tool be implemented to gain a better understanding of the current state of data quality within the Ministry. The advantage of this approach is that it minimizes the level of effort while maximizing the extent and immediacy of investment return. The proposed methodology takes full advantage of an automated approach to data quality. Details specific to each phase of the cycle are provided below.

**4.1. Identify Problems**



**Figure 3. Identify Problems**

- The goal of this stage of the analysis is the development of a catalogue of data quality issues for the data set(s) that are to be analyzed. Once the **target data** set is determined, an effort is made to capture and document the business rules that describe what good quality data is supposed to look like in that target. In other words, determine the **Data Quality Rule Set**.
- The next step involves the transformation of the collected business rules into tests that compare the actual data against the business rule standard that using a **data quality analysis and reporting tool**. The output of this effort is a **Data Quality Report Card** that identifies all of the instances where the data deviates from the stated business rule standard. Ideally, in addition to the number and percentage of the violations detected, the process should identify each row in which the violation occurs through a series of drilldown **reports**. The identification of each row can be beneficial in later analysis to track the errors back to their source or as a trigger to implement corrective updates to the data.

- It is important to ascertain if the detected violation is an actual problem. The data may be correct but the business rule has been incompletely or inaccurately stated. Defining business rules can be an iterative **review** process whereby a rule is stated and then tested, and the violations analyzed to be certain that they are true violations. Sometimes there are exceptions to a rule that must be taken into consideration. This iterative process allows organizations to accurately pinpoint data quality errors.

## 4.2. Analyze Solutions

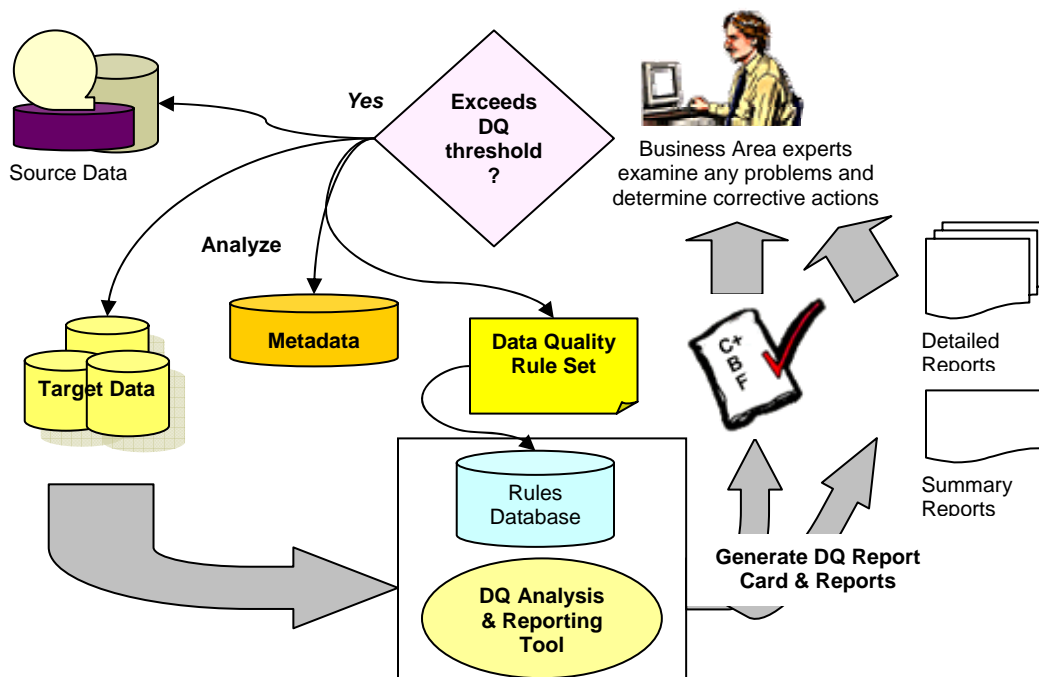


Figure 4. Analyze Solutions

- Once an error is discovered, the first consideration is whether the error is significant enough (i.e. exceeded a defined **threshold**) to warrant investigation into its source. What is the impact on the business? Are there other more pressing quality issues that need to be examined first?
- In order to track an error back to its source, it is necessary to understand where the data originates, the processes that manipulate it, and the intermediate files or tables where it may reside. Determining where a data value error is introduced is a matter of testing it in each of its upstream processes to determine where the error first appears in the **source data**.
- Once the source and reason for the error are understood, the action(s) that must be taken to correct it for the future may be considered. While corrective action can probably be devised in most situations, there may be some cases where it is simply not economically feasible to correct the problem and a different solution will have to be devised.

Some possible avenues to explore to clean up data quality issues are:

- Improve hardware, software and communications environments
- Improve and implement procedures, policies, processes
- Train data-entry personnel as required
- Implement data stewardship programs
- Use a data extraction and load process to clean up the data, if possible

### 4.3. Implement changes

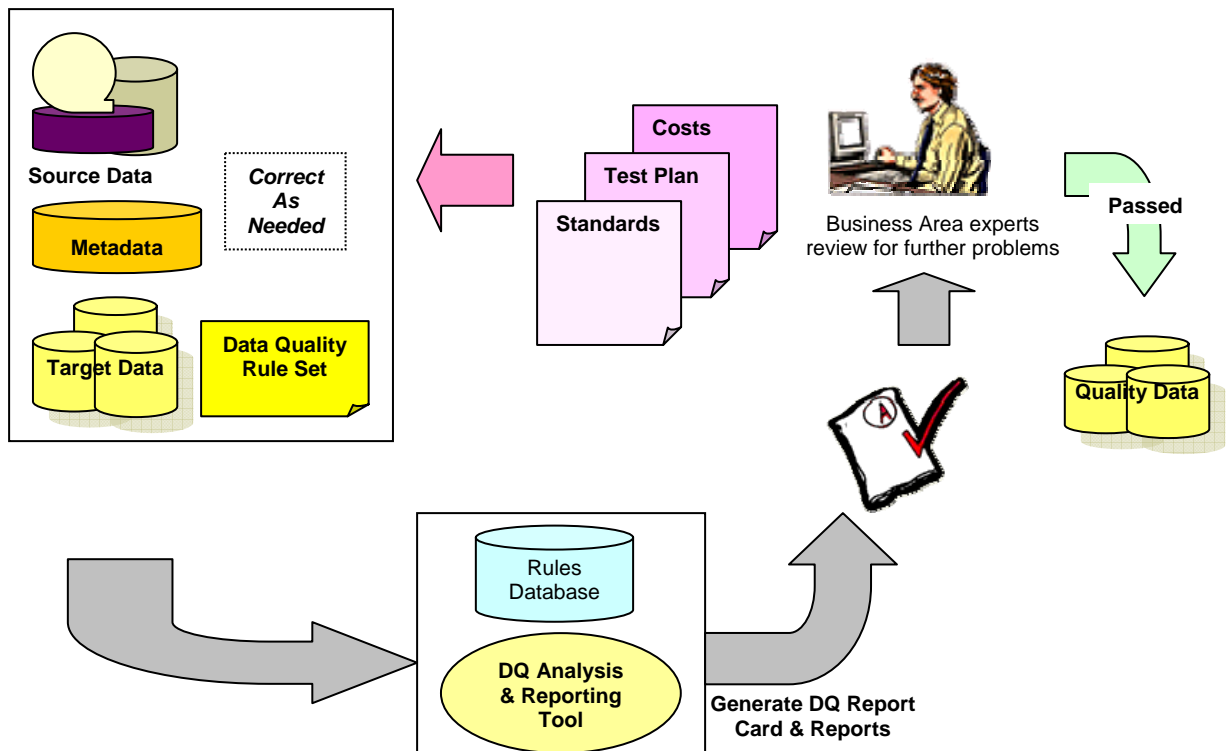


Figure 5. Implement Changes

- Corrections may be implemented through either automated or manual processes depending on the nature of the error.
- The changes decided upon in the previous step need to be implemented with a judicious eye to ensure that correcting one error doesn't inadvertently cause another. While not necessarily causing another error, it is possible that correction of one issue will highlight other issues that weren't detected in the first iteration.
- Testing and monitoring of costs as appropriate is applied during this step. Test plans and costs are updated as necessary.

#### 4.4. Monitor results

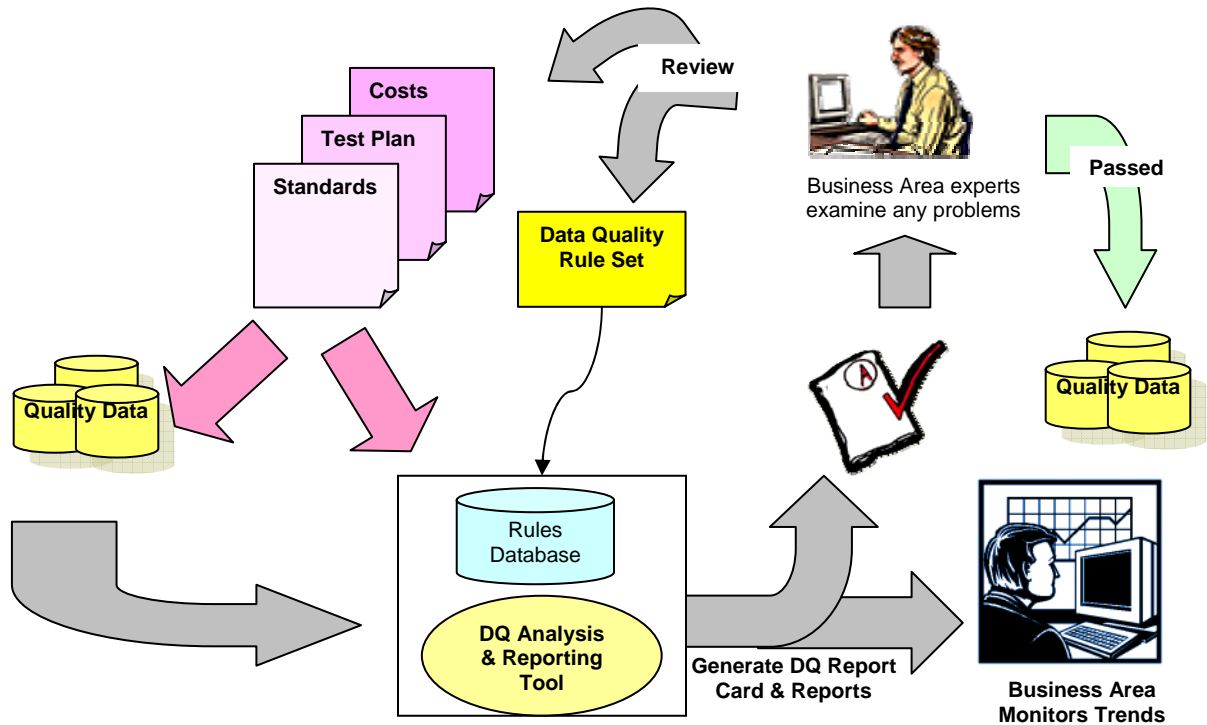


Figure 6. Monitor Results

- Once implemented, the data should be run against the data quality **standards** on a scheduled basis to ensure that the changes have had the desired results, that there hasn't been any backsliding and as an early warning of possible other data issues that may arise. Initially, the monitor needs to be run on a fairly frequent basis as changes are made to the processes and procedures. However, over time, as the data stabilizes, that frequency may be reduced. Regardless of the frequency, it is important to track the data quality **trends** over time. This process can be greatly facilitated by the use of an automated tool that can perform this monitoring and trending without significant intervention.
- Data Custodians might consider publishing the results of data quality monitoring to show high levels of quality and/or improvement over time.

#### 4.5. Relationship to Popular Quality Standards

The methodology outlined above shares fundamental characteristics with other defined data quality improvement methodologies such as Six Sigma, Total Information Quality Management, and ISO 9000:2000. A high-level summary of some representative information/data quality management methodologies is provided in Table 3. Greater detail on these methodologies plus others such as House of Quality and Kaizen are given in Appendix E.

MoFR Methodology	Six Sigma DMAIC	TIQM Processes & Steps	AIMQ Methodology	PDCA / Kaizen	Data Quality Practice	ISO 9000:2000
<b>Identify Problems</b> <ul style="list-style-type: none"> <li>Define business rules</li> <li>Report Card review</li> </ul>	<ul style="list-style-type: none"> <li>Identify the problem</li> <li>Define requirements</li> <li>Analyze priorities</li> </ul>	<ul style="list-style-type: none"> <li>Define the project</li> <li>Plan the objectives</li> <li>Identify impact</li> </ul>	<ul style="list-style-type: none"> <li>Develop questionnaire</li> </ul>	<ul style="list-style-type: none"> <li><b>Plan</b></li> <li>Identify the problem</li> <li>Set a goal</li> <li>Establish governance</li> </ul>	<ul style="list-style-type: none"> <li>Identify the problem</li> <li>Define requirements</li> </ul>	<ul style="list-style-type: none"> <li>Define requirements</li> <li>Establish quality policy</li> </ul>
<b>Analyze Solutions</b> <ul style="list-style-type: none"> <li>Analyze problems</li> <li>Determine corrections</li> </ul>	<ul style="list-style-type: none"> <li>Plan</li> <li>Measure performance</li> <li>Develop baseline</li> <li>Identify root causes</li> </ul>	<ul style="list-style-type: none"> <li>Assess Data &amp; Information Quality</li> <li>Assess costs</li> <li>Identify root causes</li> </ul>	<ul style="list-style-type: none"> <li>Assess conformance to specifications</li> <li>Do gap analysis</li> <li>Analyze gap analysis</li> </ul>	<ul style="list-style-type: none"> <li>Plan</li> <li>Identify and analyze processes</li> <li>Plan</li> <li>Identify root causes</li> </ul>	<ul style="list-style-type: none"> <li>Map the Information Chain</li> <li>Establish data quality scorecard</li> <li>Assess costs</li> <li>Identify root causes</li> <li>Build project team</li> <li>Build vs buy</li> <li>Define DQ rules</li> </ul>	<ul style="list-style-type: none"> <li>Plan objectives</li> <li>Plan &amp; develop realization processes</li> <li>Document system</li> <li>Control development</li> <li>Identify personnel</li> <li>Identify infrastructure</li> <li>Provide quality environment</li> <li>Control purchasing</li> </ul>
<b>Implement Changes</b> <ul style="list-style-type: none"> <li>Automated correction</li> <li>Bus expert correction</li> </ul>	<ul style="list-style-type: none"> <li>Implement solutions</li> <li>Test solutions</li> <li>Standardize solutions</li> </ul>	<ul style="list-style-type: none"> <li>Plan improvement</li> <li>Implement improvements</li> <li>Check impact</li> <li>Measure costs</li> <li>Standardize improvements</li> </ul>	<ul style="list-style-type: none"> <li>Test solutions with gap analysis</li> </ul>	<ul style="list-style-type: none"> <li><b>Do</b></li> <li>Develop solution</li> <li>Implement solution</li> <li><b>Check</b></li> <li>Check solution</li> <li>Standardize solution</li> </ul>	<ul style="list-style-type: none"> <li>Execute improvement</li> </ul>	<ul style="list-style-type: none"> <li>Perform remedial processes</li> <li>Establish quality system</li> <li>Develop management system</li> <li>Monitor and measure</li> </ul>
<b>Monitor Results</b> <ul style="list-style-type: none"> <li>Scheduled reporting</li> <li>Analyze trends</li> <li>Report Card review</li> </ul>	<ul style="list-style-type: none"> <li>Establish standard to maintain performance</li> <li>Correct problems as needed</li> <li>Celebrate success</li> </ul>	<ul style="list-style-type: none"> <li>Standardize quality improvements</li> <li>Report on improvements</li> <li>Improve process</li> <li>Document</li> </ul>	<ul style="list-style-type: none"> <li>Measure against others or best practices with gap analysis</li> <li>Use results to continually improve</li> </ul>	<ul style="list-style-type: none"> <li><b>Act</b></li> <li>Ongoing monitoring</li> <li>Refine solutions</li> <li>Look for other opportunities</li> </ul>	<ul style="list-style-type: none"> <li>Measure improvement</li> <li>Build on success</li> </ul>	<ul style="list-style-type: none"> <li>Analyze quality information</li> <li>Improve quality management system</li> <li>Prevent potential nonconformities</li> </ul>

**Table 3. A high-level summary of some representative information/data quality management methodologies.**

#### **4.6. What to Measure?**

The dimensions of data quality discussed in section 2.2 above and detailed in Appendix B will form the core around which business rules will be defined for a particular business area. Not all of the identified dimensions can be measured easily with a data quality analysis and reporting tool. Within the attribute dimensions, a significant number require manual intervention and/or user feedback. This makes these dimensions very difficult to examine with an automated tool unless reference data or user surveys are available digitally.

This is especially true in the case of spatial attributes. Spatial data is by necessity approximate which makes it even more difficult to assess automatically. In the short-term, the best approach for spatial data is to simply check that it exists. If there is attribute data associated with a spatial

data set, then that data set should exist. The presence or absence of spatial data relative to attribute data can be covered under the attribute dimension of Completeness.

A list of the attribute dimensions and their likelihood for use in an automated tool is given in Table 5 below. More details on the dimensions and methods of measurement are provided in Appendix B (8.2) for attribute and Appendix C for spatial.

<b>Attribute Dimension</b>	<b>Measurement can be automated?</b>	<b>Keywords</b>
Accuracy	No, unless reference set is available.	“Ground truthing”, Establish tolerances
Precision	Yes	Significant digits correct
Consistency	Yes	Common definition, Mostly implemented by DBMS or DQ tool
Believability / Reliability	No, unless surveys are available digitally	Credibility, User surveys, user complaints
Completeness	Yes	Missing values, missing related values. Spatial data exists as required.
Currency / Timeliness	No, unless measuring “time stamps”	Time stamping, “Up-to-date”, may require “Ground truthing” Possibly measured by business rule against capture date
Consistent representation	No.	Adherence to defined standard or style
Appropriateness	No, unless surveys are available digitally	Match user expectations, User surveys, user complaints
Accessibility	No, unless surveys are available digitally	Ease of access, , User surveys, user complaints Possibly capture access attempts/patterns.
Availability	No, unless surveys are available digitally	Available when needed, User surveys, user complaints
Metadata	No, unless surveys are available digitally or metadata system exists.	Metadata policies, User surveys, user complaints Possibly if reference to metadata system available.

**Table 4. Attribute dimension automations.**

## 5. The MoFR Data Quality Management Roles

### 5.1. Current Management Processes at MoFR

Data management roles and responsibilities within MoFR generally follow those outlined by the Data Administration Forum<sup>13</sup>. The roles that are most involved with the management of data quality are listed below.

#### ○ Data Custodian

Senior Manager for a business area responsible for data requirements, standards, access rules, business training, etc. Defines the business value, scope, standards and services of the organization's data within the context of their mandate.

- The Data Custodian is usually at the Director or Program Manager level.
- **Responsibility**
  - responsible for issues related to the collection, storage, protection, and delivery of their data, ensuring it meets the business needs of the organization
  - fulfilling the legislated responsibility or program mandate of ensuring data quality, completeness, and integrity through the management of its creation and maintenance
  - allocating resources in order to meet data needs
  - ensuring that the **Data Manager** defines the business standards for their data, the level of quality of the data, and that the data source is documented and the metadata published
  - developing a data management plan \*
  - ensuring adherence to government and Ministry standards \*
  - ensuring the value of data is maximized through sharing \*
- **Contact when**
  - a new business need for data is identified
  - further data services are required to meet new business needs
  - additional data may be encompassed within their business scope
  - operational, business, or data definition issues cannot be resolved
  - defined and committed services are not being supplied
  - data sharing issues arise
  - compliance issues arise with respect to government or Ministry standards
  - data management planning is required

#### ○ Data Manager

Business expert with detailed knowledge of the data structure, content, and appropriate use of the business information.

- The **Data Manager** is usually at the Program Manager level.
- **Responsibility**

---

<sup>13</sup> <http://www.cio.gov.bc.ca/other/daf/DMRolesRespV1.pdf>

- ensuring data meets business needs
- ensuring the delivery of defined services on an operational level
- ensuring the protection of data is commensurate with its value, ensuring efficient supply of services, and determining the level of quality of the data
- acting as the primary contact for business metadata
- defining and managing the acquisition and disposition of data
- facilitating the development of a Ministry wide data management strategy
- **Contact when**
  - data access, within the scope of the service, is required
  - difficulties are encountered in data access
  - data errors are perceived
  - further detailed information about their data is required

#### ○ **Data Standards Manager (Discipline Authority)**

Business expert or specialist who fully understands the business relevance of the data standards within their scope of work. They must actively use their knowledge in support of the broad scope of business and established data standards. Interprets the meaning and appropriate use of detailed data standards to meet organizational needs. A primary resource for the **Data Manager**.

- The **Data Standards Manager** is usually at the senior scientific, analytical, or research level.
- **Responsibility**
  - interpreting and defining of data within the scope of their expertise and according to the business need
  - providing expert understanding of the subject area and business definitions
  - providing guidance in the appropriate use of data and in the specification of its accuracy possibly through the definition of business views
- **Contact when**
  - understanding of data is not clear from definitions and specifications
  - interpretation of data, outside of existing specifications, is required

#### ○ **Data Usage Contact**

Technical database resource or sophisticated business user who understands the business data and how it has been physically implemented. Manipulates and queries physical database content to support operational information needs. May also define user views for repeated queries. A primary resource for the **Data Manager**.

- The **Data Usage Contact** is usually at the intermediate business or technical level.
- **Responsibility**
  - querying data content in order use data effectively according to standards

- *Contact when*
  - business-related database queries are required

## **5.2. Overall Ministry Approach**

The overall Ministry approach to data quality improvement is to first determine the level of data quality for any business area, and then for Ministry business areas to make decisions about the necessity, costs and benefits of improving their data quality.

To determine levels of data quality requires analysis tools and business rules to analyse and report on quality. To improve on data quality requires an improvement methodology, and requires management, financial and technical support.

- **Corporate Analysis Tools**

For analysis and reporting on data quality, an evaluation of commercial off-the-shelf vendor tools used to assess data quality is required. The objective of this evaluation will be to purchase a data quality analysis and reporting toolset to meet Ministry needs, and to customize it for specific MoFR requirements if necessary.

The goal is to put data quality analysis and reporting infrastructure in place that can be easily used by any Ministry business area to assess data quality.

- **Analysis Projects**

Business areas desiring greater knowledge of their data quality will supply business rules associated with the data. These rules will be translated into the syntax of the data quality analysis and reporting tool and will be used to analyse data for compliance producing both executive level quality report cards and detailed level reporting which can be used for data error correction.

An example of the amount of effort to set up data quality reporting within a data quality tool comes from a project to create business rules and set up data quality reporting for a state Department of Education in the United States. The translation of approximately 100 simple to very complex business into the Center for Data Quality data quality analysis tool Certify took approximately 1 week.

- **Data Quality Improvement Project Support**

Once reporting on data quality is available, business areas may look to the Information Management Group (IMG) for guidance, leadership and as a centre for Ministry-wide standards on data quality. IMG's role will be to:

- oversee the level of quality for the Ministry as a whole,
- offer advice on data quality best practices, and
- maintain and update the Ministry wide standards and the data quality framework.

### 5.3. Long-Term Management Goals

The realities of setting budgetary priorities limit immediate management goals, but over the long-term it will be important to focus seven basic components that will help to ensure any data quality improvement projects are successful. These seven components are:

1. Top Management Sponsorship Is Essential	Not only does top management have to say that data quality is important, it must demonstrate that commitment by funding the efforts required to ensure it. It means taking the extra time up front and during development to ensure the quality later. There must be demonstrable evidence of a constancy of this commitment to improving data quality over the long term.
2. True Data Quality Is An All-Pervasive Mindset	Beginning with the original recognition of the need for data, through all phases of analysis, design, development, capture, to archival and/or disposal, it is essential that all phases of the data lifecycle be approached with an eye on ensuring the quality of the data. Data is a critical resource of the organization and must be treated that way.
3. It is Easier To Engineer In Quality Than Retrofit It Later	Although it is not always easy to implement the changes necessary to improve the data without massive overhaul of legacy databases, some improvement is possible in every environment. A far easier approach is to follow the total quality management principles right from the design phase to engineer in the ability to maintain quality data. A quality conscious design and implementation of new software will enhance the ability to gather and manage quality data.
4. Legacy Systems Exist	Data in legacy systems is often suspect. However, even if the system is old and is “scheduled to be replaced sometime in the future”, there may be significant benefit from cleaning up the data.
5. Each Data Element Must Be “Owned”	The Data Custodians are a conduit to maximize communication between the requirements of the business community and the support givers of technology. And it is they who assume responsibility for certifying to the rest of the organization that the information in their piece of the corporate information set is maintained in a state of consistent high quality.
6. Data Quality Cannot Be Achieved With A Part-Time Effort	A function that is performed only on a part-time basis is viewed as only marginally important to the organization. An organization that seriously wants to begin managing the quality of their data as if it really is a strategic corporate resource must devote the resources to make it happen. There must be a readily recognized constancy of purpose displayed on the part of management that the quality of data is vital.
7. Data Quality Is Not A One-Time Effort	Data quality should be a continuous improvement process. That includes celebrating data quality improvement successes and improving other areas where necessary. Part

	of the improvement process should be tracking errors back to their source to attempt true corrective action there. It is important to ensure that errors don't recur. The bar on data quality should be continually raised.
--	---

## 6. Data Quality Management Tools

### 6.1. Existing Tools

The most powerful tool the Ministry has to direct, promote and guide data quality is its Data Custodian practices. Data Custodian policies, procedures and standards provide significant direction and enable sound data management practices. Each Data Custodian's ability to apply resources and influence data management practices through the Information Governance Council can provide much needed leadership in the move towards a culture of data quality.

Another powerful tool MoFR has to manage data quality is its data administration practices. Setting Ministry wide information management standards and ensuring data is designed and modelled from a corporate perspective greatly increases the clarity and reuse of data. Clarity and non-duplication are key elements in data quality.

The Ministry also has some automated tools such as its data modelling repository and its corporate databases, and some data quality analysis spreadsheets and queries in place within different business areas, which assist in defining standards for data, applying data integrity rules, and analysing data quality.

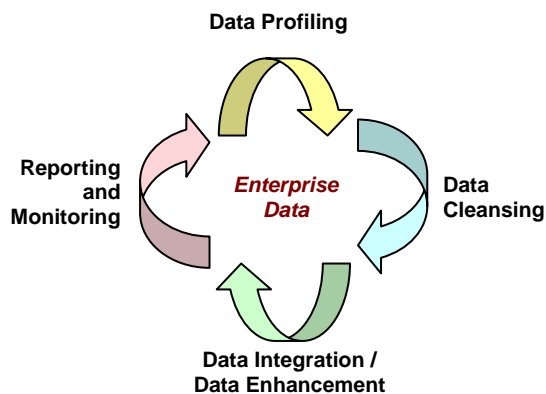
To date no Ministry-wide approach to automated analysis of data quality has been put in place.

### 6.2. Required Tools / Industry Standard Tools

Data quality analysis of data requires running large numbers of tests to validate data against the business rules. While it is possible to use SQL queries or custom written code to perform these tests, a more effective approach is the use of a software tool that is specifically focused on data quality and the various aspects that make up data quality. A tool that guides and supports the development and maintenance of tests, facilitates their efficient running and reports analysis results in a concise and meaningful fashion can significantly improve MoFR knowledge of data quality.

There are numerous products that are purported to focus on data quality<sup>14</sup>. Many of the tools take an integrated approach to data quality and break down the approach into at least four major areas as follows:

1. **Data Profiling:** Identifying data quality issues in a variety of ways, including
  - Basic statistics, frequencies, ranges and outliers
  - Identification of duplicates
  - Discovery and validation of data patterns and formats
  - Numeric range analysis
  - Identification and validation of primary/foreign key relationships across data sources
  - Validate data specific business rules within a single business area or across data sources
2. **Data Cleansing:** Correcting the data based on secondary sources, and standardizing the data to ensure a consistent format throughout all records. Process include:
  - Plan and prioritize a data correction initiative
  - Parse data into atomic components
  - Standardize and normalize data
  - Correct data values to identified standards
  - Verify and validate data accuracy
3. **Data Integration / Data Enhancement:** Merge, link and consolidate information from a variety of sources. Also generate or append additional data from other internal or external data sources. For example, merging and/or appending attribute data with spatial data.
4. **Reporting and Monitoring:** Reporting within the quality process, and monitoring data quality issues over time. For example, monitoring data quality may include:
  - Identifying trends in data quality
  - Providing alerts of violations in established data quality and business rules
  - Detecting variances from cyclical runs



**Figure 7. Data Quality Assessment**

The data quality assessment tool required by the Ministry of Forests and Range may not require all four areas described above. However, some of the elements represented in particular by the Data Profiling, and Reporting and Monitoring areas will be required. Data cleansing will also be

---

<sup>14</sup> For a list of “Information Quality” products see: <http://www.infoimpact.com/iqproducts.cfm>

highly desirable. A list of some of the basic tool elements is given in Table 5 below. An indication of the relevance of these elements to MoFR is given below.

<b>Tool Element</b>	<b>Description</b>	<b>Priority</b>
Processing Options (Data Profiling)	Connects to multiple disparate data sources simultaneously <ul style="list-style-type: none"> <li>○ Cross-data comparisons</li> <li>○ Single project organization of analysis tasks</li> <li>○ Easy viewing of distributed data sources simultaneously</li> </ul>	Medium
	Connects directly to data sources <ul style="list-style-type: none"> <li>○ Provides immediate and real-time access to dynamic data</li> <li>○ Eliminates loading into intermediate files</li> </ul>	Medium
	Supports a variety of input and output options	Medium
	Supports the definition of complex business rules	High
Testing (Data Profiling)	Automated data quality testing <ul style="list-style-type: none"> <li>○ Data Profile — includes summary profile, frequency distributions, uniqueness, completeness and duplicate columns checks</li> <li>○ Column Validation – tests for formats (e.g. phone, SIN, e-mail, etc), occurrence, patterns, pattern recognition, range, specific values</li> <li>○ Structural Integrity – tests for unique primary and foreign keys, and orphan records and ensures proper linkage between data sets</li> <li>○ Business Rules Compliance – checks compliance of data across columns or across data sets to ensure information is meeting the business rules e.g. meets specific business rule defined values or ranges</li> </ul>	High
Data Updates (Data Cleansing)	<ul style="list-style-type: none"> <li>○ Allows drill down connection to detailed data for manual correction</li> <li>○ Allows automatic correction of data to defined business rule values</li> </ul>	Medium
Presentation (Reporting)	<ul style="list-style-type: none"> <li>○ Provides a data quality “report card” <ul style="list-style-type: none"> <li>▪ Summary level: Executive level reporting. High level summaries</li> <li>▪ Detail level: Data management level reporting. Detailed analysis of data quality issues.</li> </ul> </li> </ul>	High
	○ Saves multiple, historical analysis results for trend analysis and query comparisons	Medium
	○ Outputs reports in HTML, MS Excel, XML, and other formats (as required)	Medium
	○ Emails reports for dynamic sharing of analysis results with stakeholders	Low
	○ Displays reports in pie, bar, graph, and bell curve formats for clear visual representation of data issues	Medium
	○ Provides a dashboard option	Medium
	○ Web accessible	Medium
Scheduling Option (Monitoring)	○ Collects disparate analysis tasks into a batch for building analysis for specific rules and data sets	Medium
	○ Schedules single tasks or batches to run on regular basis for automated analysis, trend building, and regression testing	High
	○ Notifies via email when scheduled task has completed	Low
	○ Allows results and reports to be exported to a centralized location	Low
	○ Issues an alerts if any of the data exceeds a threshold (i.e., breaks a business rule)	Medium
Spatial Option	Matching existence of spatial data for related attribute data.	Low

**Table 5. Automated Data Quality Tool Requirements**

The Data Profiling and Reporting options are the most immediately critical to the needs of MoFR. Reporting options at the management summary level are essential as are the more detailed

analysis tools for use by the business area experts. Some of the more critical aspects of each of these are given below.

- ***Executive/Management Level Tools***  
REPORT CARD

A Data Quality Report Card presents the results of an observation of the quality of data when compared against the business rules that define what good quality data should be. It presents in a straightforward manner summary counts and percentages of the instances in which violations of the business rules were detected.

#### TREND ANALYSIS

As the clean-up process continues, it is important to monitor the progress of that effort over time through automatic trend reporting. Where a Data Quality Report Card presents the results of a single observation of the data quality, a Trend Report presents those same statistics over a period of time. This trending tracks the improvement in the quality of the data over time to be certain that the implemented changes are having the desired effects, that there hasn't been any regression and as an early warning of other problems that may be cropping up.

#### GAP ANALYSIS

Gap Analysis Reports help to highlight the degree of separation between the standards as expressed in the data quality thresholds and the current levels of quality as disclosed in Data Quality Report Cards. This separation can help management to understand the levels of effort required to be able to achieve that threshold. It can also help to highlight the amount of progress that efforts are producing toward that goal as the gap narrows over time.

- ***Detailed Analysis Tools***

#### SPECIFIC PROBLEMS / AREAS FOR CORRECTION / IMPROVEMENT

When analyzing specific incidences of violations of business rules, it is extremely helpful to have a tool which will allow creation of customizable result tables that include all of the rows in which the violation was detected. These results sets support the drill-down analysis that is typically required to gain an understanding of the error source and to manually correct the error. When examining the result set to discern error patterns, the ability to create on-the-fly frequency distributions on any one or more columns of the result set can be an effective analysis tool.

#### CHARTS, REPORTS, INDEXES, METRICS

There are also a variety of charts and graphical displays in which data can be presented in order to facilitate understanding of issues that are used in conjunction with data quality improvement. Among these techniques are:

- **Pareto Charts** – used to graphically summarize and display the relative importance in the distribution of results of an observed event. For example, one might wish to graph the most common invalid values found in a column.
- **Control Charts** – used to track one or more data points over time as a means of determining whether any trends are apparent, similar to trend charts
- **Metrics** -- some data quality analysis tools, such as IBM's Web Sphere Audit Stage allow metric values to be assigned to the results of data quality tests in an attempt to portray the relative importance of various violations
- **Data Flow Diagrams, Information Chain Diagrams** – when tracking the flow of data elements upstream to their source, these diagramming techniques may be beneficial to the data quality analyst

### **6.3. Attaining Sustainable Data Quality Levels**

#### ESTABLISHING DATA QUALITY THRESHOLDS

Data quality thresholds are established as a result of the first analysis project performed on a data set. The initial observation of compliance to the business rules will give the first quantified indication of whether there are problems with the data and how severe those problems are. This should prompt efforts to analyze solutions, implement changes and continue monitoring the effects on the data. A side benefit of this analysis is the growing understanding by the data quality analyst of what is possible in terms of the attainable levels of quality. Ideally, all of the thresholds should be set so that the business is challenged to improve the data over time rather than accepting the status quo.

There are some data elements where anything less than 100% completeness and accuracy is unsatisfactory, perhaps due to legal issues, while others can have a lower threshold. Like everything else in business, it becomes a matter of balancing the cost to achieve and maintain a level of data quality against the business need for quality. Some of the factors that should be considered are:

- How frequently does the Ministry rely on this piece of data?
- Are there legal ramifications?
- Are business decisions made relying on this data element?
- If decisions are made, are they based on individual occurrences or in aggregate?
- If in aggregate, how many individual values go into the aggregation?
- Is the source of the data under Ministry control?
- Can changes be cost-effectively applied?
- If the source is external, can leverage be applied?
- Can the legislature help to bring pressure?
- Do we have the resources to improve the data?
- Does the data element get uploaded to any data warehouses?
- Are the answers to any of the above questions more critical in the warehouse?
- Is the data too dated to be correctable?
- 

#### WHEN 100% IS NOT REQUIRED

There are instances in business when the cost of cleaning up and maintaining data at 100% levels of quality are not justifiable in terms of the business need. It is not unusual that the diligent application of clean-up efforts can improve data to within 95% compliance with data quality standards only to

find that the business need does not justify the time, manpower or monetary resources to improve the data further. Some examples of data elements where less than perfect levels of quality may not be required are:

- Historical information that is not used in current decision making
- Non-mandatory data elements
- Elements that are more nice-to-have than mission-critical
- Elements whose business purpose doesn't justify higher quality

## COMPROMISE

In the final analysis, data quality thresholds should be a documented consensus of what is achievable and what is required by the business. It may require that all of the stakeholders who utilize the data be brought together to agree on minimum levels of quality. In some instances, it may be beneficial to set an acceptable threshold for legacy data and a second, higher threshold for all new data for the same data element, in essence saying that there isn't anything that can be done for what has been previously entered, but it is possible to improve in the future.

Data standards set by the Data Custodian become the data quality target for the organization. The current compliance statistics should be monitored and published on a scheduled basis to all stakeholders and senior management.

### ***6.3.1. Implementing Data Quality Certification***

In order to be truly effective, the goal of any data quality initiative must extend beyond basic validation and correction. Fixing errors that are detected is only the start of the effort. Each business area must continuously and pro-actively raise the bar for data quality if the quality of the information is to show improvement over time, not only in MoFR resident data sets, but also in each incoming data stream. Data Quality Certification is the ability to say to management and all database stakeholders that the data meets or exceeds all of the data quality thresholds that have been established. This is something to strive for, to be proud of when it is achieved, and something to celebrate throughout the organization.

- Perform the initial data quality observation
- Begin clean-up activities
- Establish data quality thresholds
- Monitor the data against those thresholds over time
- Publish the results for all stakeholder examination
- Track progress toward the goal
- Celebrate the success when thresholds are achieved

### ***6.3.2. International Standards***

International standards around quality focus more on the management of quality rather than the specifics of individual data dimensions. Two major organizations that provide standards and related certifications around the management of quality are ISO and NQI.

- **ISO 9000**

ISO (International Organization for Standardization) is the world's largest developer of standards. ISO 9000 has become an international reference for quality management. Specifically what an organization does to fulfil:

- the customer's quality requirements, and
- applicable regulatory requirements, while aiming to
- enhance customer satisfaction, and
- achieve continual improvement of its performance in pursuit of these objectives.<sup>15</sup>

The advantage of establishing an ISO 9000 process is that it provides a “tool” to achieve repeatable levels of quality. This means that the processes that create data models, databases, applications, and processes that create or update information should be documented to achieve repeatable quality. Another advantage of ISO certification is that it is internationally recognized and provides an important external validation for an organizations quality management practices.<sup>16</sup>

- **NQI**

The National Quality Institute (NQI) is an independent, not-for-profit organization that is a leading authority in Canada on workplace excellence based on quality systems.<sup>17</sup> NQI offers an implementation criterion called the Progressive Excellence Program (PEP) which provides the framework and strategic context for ISO and other systems so that an organization can incorporate other improvement initiatives into an integrated management system. Most importantly, NQI-PEP is based on NQI Criteria (Canadian Framework for Business Excellence & Canadian Quality Criteria for Public Sector Excellence), and designed specifically to assist organizations develop a planned target driven approach. The Canadian Quality Criteria for the Public Sector serves as a framework for effective public service organizations and agencies at all levels.<sup>18</sup>

## **6.4. Organizational Impact**

The goal of this effort is to understand and improve data quality where desired and cost justified. It is also to nurture a culture change over time so that employees, contractors, and licensees view data quality as key to Ministry success, and the Ministry of Forests and Range projects an image of quality to government, business partners and the public.

- **Management Considerations**

---

<sup>15</sup> <http://www.iso.org/iso/en/iso9000-14000/understand/inbrief.html>

<sup>16</sup> Other advantages, based on the eight quality management principles defined in ISO 9000:2000, are available at: <http://www.iso.org/iso/en/iso9000-14000/understand/qmp.html>

<sup>17</sup> <http://www.nqi.ca/about/>

<sup>18</sup> For additional details specific to MoFR see: Chen, T. (2005) "FREP Quality Assurance Framework: Background Paper." Ministry of Forests and Range (Forest Practices Branch), Ministry of Environment, Integrated Land Management Agency: Internal Report.

This type of cultural change does not happen overnight or without cost. It requires a long-term commitment to the effort. Once data quality performance measures are adopted throughout, the Ministry will avoid repetition of costly data cleanups and will enjoy sound business decisions based on quality data. Management must never lose the opportunity to champion the data quality improvement effort.

### ***6.5. Beyond data quality management tools***

The ultimate objective of the framework and the data quality program is to encourage a constant striving for quality data within the Ministry. A data quality analysis and reporting tool will go a long way in helping to establish a culture of quality. However, for those dimensions that do not readily lend themselves to an automated approach (e.g. believability, currency, appropriateness) more creative solutions will be required where deemed important by a business area. Solving those data quality issues highlighted by the automated tool will help address dimensions such as believability or reliability as a by product.

## 7. Appendix A – References

Ref No	Reference
1	Aaldeers, H. (2002) "The Registration of Quality in a GIS" IN Shi, W., P.. Fisher, & M.F. Goodchild (Eds) (2002) "Spatial Data Quality." Taylor and Francis. Pg 186-199
2	Becker Associates (2005) The House of Quality. <a href="http://www.becker-associates.com/house_of_quality.HTM">http://www.becker-associates.com/house_of_quality.HTM</a>
3	Chen, T. (2005) FREP Quality Assurance Framework: Background Paper. Ministry of Forests and Range (Forest Practices Branch), Ministry of Environment, Integrated Land Management Agency: Internal Report.
4	Crow, K. (2000) Performing QFD Step by Step. DRM Associates. ( <a href="http://www.isixsigma.com/offsite.asp?A=Fr&amp;Url=http://www.npd-solutions.com/qfdsteps.html">http://www.isixsigma.com/offsite.asp?A=Fr&amp;Url=http://www.npd-solutions.com/qfdsteps.html</a> )
5	English, L. (1999) Improving Data Warehouse and Business Information Quality.: Methods for reducing costs and increasing profits. John Wiley & Sons, Inc 518 pg.
6	English, L. (2003) Total Information Quality Management: A Complete Methodology for IQ Management. DM Review feature 2003-09
7	English, L. (2004) Six Sigma and Total Information Quality Management (TIQM). Information Impact Internation, Inc. <a href="http://www.infoimpact.com">http://www.infoimpact.com</a>
8	Frank, S. & D. Brunson (2006) "The Partnership of Six Sigma and Data Certification. Business Intelligence Network." ( <a href="http://www.b-eye-network.com/newsletters/ben/2263">http://www.b-eye-network.com/newsletters/ben/2263</a> )
9	Goyal, N & L. Bhatia (2006) Improving Financial Services Through TQM: A Case Study. iSixSigma, LLC. <a href="http://finance.isixsigma.com/library/content/c040127a.asp">http://finance.isixsigma.com/library/content/c040127a.asp</a>
10	Jakobsson, A. (2002) "Data Quality and Quality Management - Examples of Quality Evaluation Procedures and Quality Mangement in European National Mapping Agencies. IN Shi, W., P.. Fisher, & M.F. Goodchild (Eds) (2002) "Spatial Data Quality." Taylor and Francis. pg 216-229
11	Lee, Y.W., D.M. Strong, B.K. Kahn, & R.Y. Wang (2002) AIMQ: a methodology for information quality assessment. Information & Management 40: 133-146.
12	Loshin, D. (2001) Enterprise Knowledge Management: The Data Quality Approach. Academic Press. 493pp.
13	North Carolina Department of Environment and Natural Resources (2002) Plan-Do-Act-Check: A problem solving process. <a href="http://www.isixsigma.com/offsite.asp?A=Fr&amp;Url=http://quality.enr.state.nc.us/tools/pdca.htm">http://www.isixsigma.com/offsite.asp?A=Fr&amp;Url=http://quality.enr.state.nc.us/tools/pdca.htm</a>
14	Praxiom Research Group Ltd (2005) ISO 9001 2000 Translated into Plain English. ( <a href="http://praxiom.com/iso-9001.htm">http://praxiom.com/iso-9001.htm</a> )
15	Shi, W., P.. Fisher, & M.F. Goodchild (Eds) (2002) "Spatial Data Quality." Taylor and Francis.313pp.
16	Vitalo, R. (2005) Six Sigma and Kaizen Compared: Part 1. Vital Enterprises. <a href="http://www.vitalentusa.com/learn/6-sigma_vs_kaizen_1.php">http://www.vitalentusa.com/learn/6-sigma_vs_kaizen_1.php</a>
17	Wikipedia (2006) Kaizen. ( <a href="http://en.wikipedia.org/wiki/Kaizen">http://en.wikipedia.org/wiki/Kaizen</a> )

## 8. Appendix B – Alphabetical Listing of Attribute Data Quality Dimensions

(References are given in square brackets)

Dimension	Definition	Measured By
<b>Accessibility</b>	Accessibility refers to the degree of ease of access to information, as well as the breadth of access (whether all the information can be accessed).	Measure by answering: (1) For each data set, how easy is it to automate access? (2) Does the presentation allow for the display of all data? (3) Is the presentation in a form that allows the user to absorb what is being presented? (4) How easy is it to get authorized to access the information? (5) Are there filters in place to block unauthorized access? [12]
<b>Accuracy</b>	Accuracy refers to how closely the data value agrees with the correct or "true" value. Accuracy may also refer to non-quantitative data, such as customer names, customer addresses, customer segment categorization, product classifications and descriptions.	Accuracy is measured by comparing the given values with the identified correct source. The simplest metric is a ratio of correct values and incorrect values. A more interesting metric is the correctness ratio along with a qualification of how correct the values are using some kind of distance measurement. [12]
<b>Accuracy (to reality)</b>	See "Accuracy"	
<b>Accuracy to surrogate source</b>	Is a measure of the degree to which data agrees with data contained in an original source of data, such as a form, document, or unaltered electronic record received within the control of the organization.	The measure accuracy to surrogate source is an assessment of the percent of records whose values for a given field are accurate as compared with the values contained in that "authoritative" source of information. [5]
<b>Appropriateness</b>	Appropriateness is the dimension that categorizes how well the format and presentation of the data matches the users' needs.	To measure this dimension, must explore the history of the interaction between the user group and the designers and implementers. If there are many occurrences of user requests that result in changes to the data model or to the data presentation layer, this may indicate a low level of appropriateness. [12]
<b>Appropriate Amount</b>	Whether the amount of information is sufficient for needs.	Rate in terms of: (1) This information is of sufficient volume for our needs; (2) The amount of information does not match our needs; (3) The amount of information is not sufficient for our needs; (4) The amount of information is neither too much nor too little. [11]

Dimension	Definition	Measured By
<b>Availability</b>	Data is only useful if it is available when needed. This is especially true for managers relying on decision support systems. Often, systems are down and data is not accessible during maintenance periods or system failures.	Data availability can be measured as the ratio of the amount of time data is available to the amount of time data is needed for access. [12]
<b>Believability</b>	The degree of credibility or trustworthiness of the information.	Measure in terms of: (1) This information is believable; (2) This information is of doubtful credibility; (3) This information is trustworthy; (4) This information is credible. [11]
<b>Completeness</b>	Completeness refers to the expectation that certain attributes are expected to have assigned values in a data set. Completeness also pertains to retention requirements for historical data and to the expectation that required associated records will be present.	Completeness rules can be assigned to a data set in four levels of constraints: (1) Mandatory attributes that require a value; (2) Optional attributes, which may have a value; (3) Inapplicable attributes (e.g. maiden name for a single male) which may not have a value. (4) Mandatory associated records. [12]
<b>Concise Representation</b>	The information is formatted compactly and presented concisely.	Rate on the basis of the following: (1) This information is formatted compactly; (2) This information is presented concisely; (3) This information is presented in a compact form; (4) The representation of this information is compact and concise. [11]
<b>Concurrency of redundant or distributed data</b>	Concurrency is the information float or lag time between when data are knowable (created or changed) in one database and are also knowable in a redundant or distributed database.	The measure concurrency is an assessment of the average length of time from when records are created or updated in one database until the time the same records (or their semantic equivalent record) are propagated to another database. An alternative measure is the percent of records propagated to another database by a specified time. [5]
<b>Consistency</b>	Data consistency refers to the common definition, understanding, interpretation and calculation of a data element. Consistency constraints can be encapsulated as a set of rules that specify consistency relationships between values of attributes, either across a record or message, or along all values of a single attribute. These consistency rules can be applied to one or more dimensions of a table - or even across tables.	The simplest measure is to perform record linkage among the data sets under investigation and verify that the shared attributes in each set have the same values. The reported measure is the number of entities that do not have consistent representation across the enterprise. [12]

Dimension	Definition	Measured By
<b>Consistent Representation</b>	This dimension refers to whether instances of data are represented in a format that is consistent with the domains of values as well as consistent with other similar attribute values.	One way to measure this dimension is to see whether there is a style guide for data representation throughout the enterprise. A more granular investigation should be done to determine if there is a standard representation format associated with every base data type and domain. The next step would examine all presentations of values associated with every data type and domain and see if the representation is consistent with the standard representation. [12]
<b>Contextual clarity</b>	The relative degree to which data presentation enables the knowledge worker to understand the meaning of the data and avoid misinterpretation.	The measure contextual clarity is a subjective measure of the ease with which information as presented is understandable by the knowledge worker. Objective measures include the percent of correct actions taken as a result of presented information. [5]
<b>Correct Interpretation</b>	A good presentation provides the user with everything required for the correct interpretation of information.	If an application has an online help facility can use this to measure the correct interpretation dimension. The help facility can be augmented to count the number of times a user invokes help and log the questions the user asks. Applications without online help can still be evaluated. This is done by assessing the amount of time the application developer spends explaining, interpreting, or fixing the application front end to enhance the user's ability to correctly interpret the data. [12]
<b>Currency</b>	Currency refers to the degree to which information is current with the world that it models.	Currency can measure how "up-to-date" information is and whether it is correct despite possible time-related changes. [12]
<b>Definition conformance</b>	The consistency of meaning of the actual data values with its data definition.	Definition conformance is the degree of agreement between the "meaning" people assign to data and its "official" definition. For detailed measures see [5]
<b>Derivation integrity</b>	Derivation integrity is the correctness with which two or more pieces of information are combined to create new information.	The measure derivation integrity is an assessment of the percent of correctness of the calculations of derived data according to the derivation formula or calculation definition. [5]

Dimension	Definition	Measured By
<b>Ease of Operations</b>	The information is easy to manipulate and/or aggregate.	Rate in terms of: (1) This information is easy to manipulate to meet our needs; (2) This information is easy to aggregate; (3) This information is difficult to manipulate to meet our needs; (4) This information is difficult to aggregate; (5) This information is easy to combine with other information. [11]
<b>Equivalence of redundant or distributed data</b>	Equivalence of redundant or distributed data is the degree that data in one data collection or database is semantically equivalent to data about the same object or event in another data collection or database. Semantic equivalence means that the values are conceptually equal; in other words, they mean the same thing in both places.	The measure equivalence is an assessment of the percent of field in records within one data collection that are semantically equivalent to their corresponding fields within another data collection or database. [5]
<b>Flexibility</b>	Flexibility in presentation describes the ability of the system to adapt to changes in both the represented information and in user requirements for presentation of information.	Can measure flexibility across two axes: (1) counting the number of times users make requests for changes in the presentation and (2) measuring the difficulty in implementing a change. That is measured either in the amount of time required, the number of parties involved, or the number of files, tables, or programs that need to be modified. [12]
<b>Format Precision</b>	The presentation of an attribute's value should reflect the precision of the value based on both the internal representation and the needs of the users.	To measure this dimension, it is necessary to prioritize display values based on the importance placed on the users' ability to differentiate by degree of precision. In other words, isolate those variables to which users focus and measure how well the precision conforms to the users' expectations. [12]
<b>Free of Error</b>	The information is correct, accurate and reliable.	Rate on the basis of the following: (1) This information is correct; (2) This information is incorrect; (3) This information is accurate; (4) This information is reliable. [11]
<b>Metadata</b>	Presence of an enterprise-wide metadata framework and support policies.	To measure the metadata policy, must score the following questions: (1) Is there a metadata policy defined?, (2) Is there a metadata repository? (3) Where is metadata stored and under whose authority? Is metadata stored in a location accessible to all users? Can users browse the metadata, especially if they are integrating a new information system component?

Dimension	Definition	Measured By
<b>Null Values</b>	A check on missing data. Should have null value rules to specify whether a data field may or may not contain null values.	Count the number of empty or null attributes that are expected to have a value. More precise measurements can be collected after specifying the exact rules regarding the types of null values and measuring conformance to those rules.
<b>Interpretability</b>	The information is easy to interpret. The codes are clear. The measurement units are clear.	Rate on the basis of the following: (1) It is easy to interpret what this information means; (2) This information is difficult to interpret; (3) It is difficult to interpret the coded information; (4) This information is easily interpretable; (5) The measurement units for this information are clear. [11]
<b>Objectivity</b>	The information was objectively collected and is based on facts.	Rate on the basis of the following: (1) This information was objectively collected; (2) This information is based on facts; (3) This information is objective; (4) This information is applicable to our work. [11]
<b>Portability</b>	Portability with respect to presentation incorporates the use of standards and recognized symbolism and the ability to perform context-sensitive customization. Measured on the basis of: Subscription to data standards; Ability to internationalize; Ability to allow personalized customization; Use of known symbols and icons; Platform transparency.	Measure portability based on: (1) Subscription to data standards, (2) Ability to internationalize, (3) Ability to allow personalized customization, (4) Use of known symbols and icons, & (5) Platform transparency (does the presentation remain the same when seen from different hardware or software platforms). Grade the answers to these to provide a measurement of portability.
<b>Precision</b>	Precision is the ability of a measurement or analytical results to be consistently reproduced, or the number of significant digits to which a value has been measured or calculated. One can simultaneously be extremely precise and totally inaccurate.	The measure of precision is an assessment of the percent of records having values to the right degree of granularity for a specified field. [5]
<b>Privacy</b>	Privacy is an issue of selective display of information based on internally managed permissions. Privacy is a policy issue that may extend from the way the data are stored and encrypted to the means of transference and whether the information is allowed to be viewed in a nonsecure location.	To measure a privacy policy, want to ask the following: (1) Is there a privacy policy? (2) If there is a privacy policy in place, how well is privacy protected? (3) Are there safeguards in place to maintain privacy and confidentiality? [12]

Dimension	Definition	Measured By
<b>Redundancy</b>	Data redundancy refers to the acquisition and storage of multiple copies of equivalent data values.	In order to measure redundancy, must look to the following issues: (1) Is redundancy planned in the system or not? (2) If redundancy is planned, what is the hardware infrastructure for storage? (3) What is the policy for copy updates? Is updating performed in real time, or is synchronization performed across the enterprise at a specific time? (4) Who manages the source copy? Or are there multiple copies viewed as equals with a synchronization process coordinating all copies? (5) If redundancy is unplanned, is it undesired? (6) With unplanned redundancy, how well are the copies synchronized? (7) If redundancy is unplanned, how do multiple copies affect the efficiency of operations? (8) How do multiple copies affect data synchronization across the enterprise? [12]
<b>Relevancy</b>	The information is useful / relevant / appropriate / applicable to our work.	Rate on the basis of the following: (1) This information is useful to our work; (2) This information is relevant to our work; (3) This information is appropriate for our work; (4) This information is applicable to our work. [11]
<b>Reliability</b>	While reliability is closely related to accuracy, it is more of a relative measure of how much confidence one can place in the data values. Reliability is often used for data that is provided from external providers.	See Objectivity
<b>Representation of Null Values</b>	Must have a recognizable form for presenting null values that does not conflict with any valid values.	Should consider the following factors: (1) Are there special ways of representing null values?, (2) If user-defined null types are used, can the user distinguish between null types in the presentation?, (3) Can the user distinguish between a null value and valid default or 0 / blank values? [12]
<b>Reputation</b>	The information has a good reputation for quality. It comes from good sources.	Rate on the basis of the following: (1) This information has a poor reputation for quality; (2) This information has a good reputation; (3) This information has a reputation for quality; (4) This information comes from good sources. [11]

<b>Dimension</b>	<b>Definition</b>	<b>Measured By</b>
<b>"Rightness," or fact completeness</b>	Rightness is the characteristic of having the right kind of data with the right quality to support a given process. It is a measure of completeness of the kinds of facts required to support a process or decision.	The measure rightness is an assessment of the percent of fact types, weighted, available out of the total fact types required to support a specific process. [5]
<b>Security</b>	The information is protected against unauthorized access. Security is similar to privacy, except that privacy deals with protecting the entities being modeled by the system, whereas security protects the data themselves.	To measure security policy, want to ask: (1) Are there different storage procedures for confidential data versus nonconfidential data? (2) How does the policy enforce security constraints (such as loading secure information onto a portable data device like a laptop, PDA, or even pages and mobile telephones)?
<b>Timeliness [/Freshness]</b>	Timeliness refers to the time expectation for accessibility of information. Timeliness can be measured as the time between when the information is expected and when it is readily available for use.	Define the time criteria and constraints that are expected for the arrival of data and then measuring in the period how frequently the data are available when expected. Is also interesting to measure how late the data actually are. Examine the range of lateness over a period of time.
<b>Understandability</b>	The meaning of this information is easy to understand.	Rate on the basis of the following: (1) This information is easy to understand; (2) The meaning of this information is difficult to understand; (3) This information is easy to comprehend; (4) The meaning of this information is easy to understand.
<b>Uniqueness</b>	Uniqueness is closely related to consistency. For a data element to be consistent, it should also have a unique identity and definition.	There must be a unique definition and method of calculation for "lifetime value" and "large" customers to calculate the lifetime value of large customers. [5]
<b>Unit Cost</b>	The cost of maintaining information contributes greatly to an organization's ability to provide information-based services, information processing, as well as general overall efficiency.	To measure unit cost, must measure (1) the cost to obtain values, (2) the cost of storage, (3) the cost of processing per unit, (4) the cost to maintain levels of data quality, and (5) the cost of building data quality into an information product. [12]
<b>Usability</b>	Usability is the relative ease of use of the form of information presentation required to support the information use.	The measure usability is a subjective measure of the degree to which the information presentation is directly and efficiently usable for its purpose, such as to perform a process or support a decision. [5]

Dimension	Definition	Measured By
<b>Use of Storage</b>	Disk space may be inexpensive, but it is not unlimited. A dimension of data quality, therefore, is in evaluation of storage use. Want to investigate how effectively the storage requirements are offset by other needs, such as performance or ease of use.	Considerations include: Storage performance meeting user requirements; How well storage will scale; Are there performance glitches inherent in the architecture; Is data replication being used to good advantage; Should data model be denormalized for performance. [12]
<b>Validity or business rule conformance</b>	Validity is a measure of the degree of conformance of data values to its domain and business rules. Validity means the data value is from the correct domain of values for a field.	The measure validity is an assessment of the percent of records having values that conform to the tested business rules for a field. [5]

### 8.1. A comparison of attribute dimensions

	English [5]	Frank & Brunson [8]	Lee, Strong, Kahn & Wang [11]	Loshin [12]
<b>Intrinsic IQ</b>	Accuracy (to reality), accuracy to surrogate source, precision, nonduplication, equivalence of redundant or distributed data, concurrency of redundant or distributed data, derivation integrity	Accuracy/precision, reliability, consistency, uniqueness	Believability, reputation, objectivity, free of error	Accuracy, consistency
<b>Contextual IQ</b>	Definition conformance, completeness (of values), validity or business rule conformance, timeliness, "rightness," or fact completeness	Completeness, timeliness	Relevancy, completeness, timeliness, appropriate amount	Null values, completeness, currency/timeliness
<b>Representational IQ</b>	Contextual clarity, usability		Understandability, interpretability, concise representation, consistent representation	Appropriateness, correct interpretation, flexibility, format precision, portability, representation consistency, representation of null values
<b>Accessibility IQ</b>	Accessibility	Availability	Accessibility, ease of operations, security	Accessibility, metadata, privacy & security, redundancy, unit cost, use of storage

## 8.2. MoFR attribute data quality dimensions details

The following attribute dimensions have been identified as having relevance and applicability within MoFR. The importance or priority of each dimension will vary with the business area. Those dimensions regarded as having a high priority across the Ministry are indicated. Details for each of the dimensions follows.

Dimension (Priority)	Measurement Type	Keywords
Accuracy (High)	Mostly manual	“Ground truthing”, Establish tolerances
Precision	Automated	Significant digits correct
Consistency (High)	Automated	Common definition, Mostly implemented by DBMS or DQ tool
Believability / Reliability	Manual	Credibility, User surveys, user complaints
Completeness (High)	Automated	Missing values, missing related values
Currency / Timeliness (High)	Automated for time stamping, Otherwise manual	Time stamping, “Up-to-date”, may require “Ground truthing” Possibly measured by business rule against capture date
Consistent representation	Manual	Adherence to defined standard or style
Appropriateness	Generally manual	Match user expectations, User surveys, user complaints
Accessibility	Manual	Ease of access
Availability	Generally manual	Available when needed, User surveys, user complaints
Metadata	Generally manual	Metadata policies, User surveys, user complaints

### 8.2.1. Accuracy

#### ○ Definition

Accuracy refers to how closely the data value agrees with the correct or “true” value. There are many different sources of correct information: a database of record, a similar, corroborative set of data values, dynamically computed values, the result of manual workflow, or irate users.

#### ○ Measure By

- Ratio of incorrect data values to correct values
  - “Correct” values determined by “ground truthing”, surveys, GPS etc.
  - Need to establish tolerances
  - Complex business rules

#### ○ Automation

May be automated if “correct” values are available digitally

### 8.2.2. Precision

#### ○ Definition

Precision is the ability of a measurement or analytical results to be consistently reproduced, or the number of significant digits to which a value has been measured or calculated. One can simultaneously be extremely precise and totally inaccurate.

#### ○ **Measure By**

- The percent of records having values to the right degree of granularity for a specified field

#### ○ **Automation**

Can be automated.

### **8.2.3. Consistency**

#### ○ **Definition**

Data consistency refers to the common definition, understanding, interpretation and calculation of a data element. This consistency to a common definition also applies the implementation of a data architecture and business rules around data dependencies.

#### ○ **Measure By**

- Value Format or Structure
  - Measure the percent of records matching the specified format for that data.
  - For example, the number of feature codes matching the CCSM catalogue format of ten characters (two letters, and eight digits).
- Valid Value Set or Ranges
  - The allowed value sets of some columns are bounded by a finite list of valid values or acceptable range of values.
  - For example, topographic features using the CCSM classification must have:
    - A single letter, representing one of ten major classes.
    - A single letter, representing categories of the major classes.
    - Five digits, assigned sequentially across all categories, linked to a feature definition.
    - Three digits, used to encode attributes of the feature.
  - Measure the number of values that do not meet the determined valid value set or range.
- Primary Key Uniqueness
  - The uniqueness of the primary key is generally enforced by the database management system. However, external source data may not have a unique primary key and so may need to be tested.
- Row Uniqueness
  - This measure is about duplication of data. Within the appropriate context, a cut block or tenure should appear only once. Duplicate entries are an error.
  - A count of duplicated data will give a measure of quality for this element of the Consistency dimension.
- Referential Integrity
  - Primary key and foreign key referential integrity is generally enforced by the database management system. However, external source data may contain referential integrity errors and so should be tested.
- Dependency Constraints

- There are many instances where the values in one column are dependent upon the values in some other column or columns in the same table or in other related tables. Complex business rules may guide the relationships.
- For example, if a Forest Client has a Client Type Code of I (individual) First Name must be present. If the Client Type Code is C (corporation) First Name must be blank (null).
- Measure the number, as a percentage, of instances where values do not meet the business rule constraints.
- **Validity Table**
  - Valid value combinations from two or more columns that are used to test the validity of source data.
  - Measure the number, as a percentage, of instances where the value is not valid..
- **Cardinality**
  - Restrictions placed on the number of times that a foreign key value can appear in the value set of a child table.
  - For example, within a Forest Cover Polygon (area), a Tree Layer may only have one related record showing the species percentage for any Tree Species.
  - Measure the number of times the restriction is not met.

#### ○ **Automation**

Can be automated to a large extent by a database management system or data quality analysis and reporting tool.

### **8.2.4. Believability / Reliability**

- **Definition**  
The degree of credibility or trustworthiness of the information.
- **Measure By**
  - May be determined by survey of customers or unsolicited customer feedback.
- **Automation**  
Generally manual, but may be automated if survey results are available digitally.

### **8.2.5. Completeness**

- **Definition**  
Completeness refers to the expectation that certain attributes are expected to have assigned values in a data set. Included in this dimension are aspects that may be enforced by a database management system.
- **Measure By**
  - **Data population (includes Null Values)**
    - Data population rules dictate whether a column can contain null values, or blanks or zeroes depending upon the data type, or some other default value to represent a “unknown” condition.
    - Measure the number of missing values

- Mandatory attributes that require a value
  - Optional attributes that may have a value
- Measure number of missing related values
  - For example, if harvesting is complete for a cut block, an opening must exist.
  - Complex business rules may guide the relationships.
- Measurements of the Consistency dimension elements of Referential Integrity and Primary Key Uniqueness will also measure “Completeness” to some extent.
  - For Referential Integrity, a foreign key in a child table must exist as a primary key in the parent table.
  - For Primary Key Uniqueness, a primary key is by definition Not Null. Any missing (not complete) primary keys invalidate the definition of a primary key.
- **Automation**
  - Can be automated.

### **8.2.6. Currency / Timeliness**

#### ○ **Definition**

Currency refers to the degree to which information is current with the world that it models. Currency can measure how “up-to-date” information is and whether it is correct despite possible time-related changes.

#### ○ **Measure By**

- If data has a standard update date, can check timestamp to ensure update is complete.
  - Can measure how frequently data are available when expected.
- Can measure data against current state on the ground.
  - Data may be accurate in terms of position, or in terms of some time in the past but still not be current or timely.
  - May require ground truthing.

#### ○ **Automation**

- Can be automated
  - Can check timestamps against standard update dates.
  - Can calculate percent of data that is current or timely against most up-to-date data or ground truth data if available digitally.
  - Otherwise may require manual intervention.

### **8.2.7. Consistent Representation**

#### ○ **Definition**

This dimension refers to whether data elements or symbols are consistent with a defined standard or style. For example, are instances of the data represented in format that is consistent with the domain of values as well as consistent with other similar values?

#### ○ **Measure By**

- Examine all presentations of values associated with every data type and domain and see if the representation is consistent with the standard representation.
- For example, representation is consistent across iMap, MapView, ArcGIS and layer files.

- **Automation**
  - Measured manually

### **8.2.8. Appropriateness**

- **Definition**

Appropriateness is the dimension that categorizes how well the format and presentation of the data matches the users' needs.

- **Measure By**
  - May be determined by survey of customers or unsolicited customer feedback.

- **Automation**
  - Generally manual, but may be automated if survey results are available digitally.

### **8.2.9. Accessibility**

- **Definition**

Accessibility refers to the degree of ease of access to information, as well as the breadth of access (whether all the information can be accessed).

- **Measure By**
  - Measure by answering:
    - (1) For each data set, how easy is it to automate access?
    - (2) Does the presentation allow for the display of all data?
    - (3) Is the presentation in a form that allows the user to absorb what is being presented?
    - (4) How easy is it to get authorized to access the information?
    - (5) Are there filters in place to block unauthorized access?

- **Automation**
  - Measured manually

### **8.2.10. Availability**

- **Definition**

Data is available when it is needed.

- **Measure By**
  - Data availability can be measured as the ratio of the amount of time data is available to the amount of time data is needed for access.

- **Automation**
  - Generally manual, but may be automated if user survey results are available digitally.

## **8.2.11. Metadata**

### **○ Definition**

Presence of an enterprise-wide metadata framework and support policies.

### **○ Measure By**

- To measure the metadata policy, must score the following questions:
  - (1) Is there a metadata policy defined?,
  - (2) Is there a metadata repository?
  - (3) Where is metadata stored and under whose authority?
  - (4) Is metadata stored in a location accessible to all users?
  - (5) Can users browse the metadata, especially if they are integrating a new information system component?
  - (6) Is a data custodian assigned?

### **○ Automation**

- Generally manual, but may be automated if user survey results are available digitally.

## 9. Appendix C – Alphabetical Listing of Spatial Data Quality Dimensions

Name	Definition	Further details / Measured by
<b>Absolute or external accuracy</b>	Closeness of reported coordinate values to values accepted as or being true.	The accuracy measured (in both the horizontal (x,y) and vertical (z) dimensions) against the spatial reference system.
<b>Accuracy of a time measurement</b>	Correctness of the temporal references of an item (reporting of error in time measurement)	
<b>Classification correctness</b>	Comparison of classes assigned to features or their attributes to a universe of discourse.	
<b>Cloud cover</b>	Area of data set obstructed by clouds.	Expressed as a percentage of spatial extent.
<b>Commission</b>	Excess data present in the dataset, as described by scope.	
<b>Completeness</b>	Presence and absence of features, their attributes and their relationships.	Indicates the estimation of errors of omission and commission which can be expressed by percentages of missing or over-complete data in the data set relative to the specification.
<b>Conceptual consistency</b>	Adherence to rules of conceptual schema	
<b>Data quality element information</b>	Quantitative descriptions of measurable aspects of the performance of occurrences in the data set for an intended use.	
<b>Domain consistency</b>	Adherence of values to the value domains	
<b>Format consistency</b>	Degree to which data is stored in accordance with the physical structure of the dataset, as described by the scope	
<b>Gridded data position accuracy</b>	Closeness of gridded data position values to values accepted as or being true.	Grid-points and pixels are defined by a couple of integers representing the row and column numbers in a grid.
<b>Homogeneity</b>	Describes how well the quality information is applicable to all occurrences of entities in the data set.	
<b>Lineage</b>	Information about the events or source data used in constructing the data specified by the scope or lack of knowledge about lineage.	The description of the processing history which each occurrence of an entity has undergone since its original creation. For each process, a statement has to be given describing the processing the entity has undergone (including details of the methods used and possibly references to documents containing actual algorithms applied), who performed the process, when it was performed, and why.

Name	Definition	Further details / Measured by
<b>Logical consistency</b>	Degree of adherence to logical rules of data structure, attribution and relationships (data structure can be conceptual, logical or physical)	Consistency includes static and dynamic consistency. Static consistency describes the validation of data and constraints of data relations, while dynamic consistency describes the validation of the processes.
<b>Metric accuracy</b>	See Positional Accuracy	
<b>Non-quantitative attribute correctness</b>	Correctness of non-quantitative attributes.	Possible metrics include the probability of a correct classification or misclassification, and the probability of correctly assigning alternative values.
<b>Omission</b>	Data absent from the dataset, as described by scope.	
<b>Positional accuracy</b>	Accuracy of the position of features.	
<b>Process step information</b>	Information about an event in the creation process for the data specified by the scope	Includes a description of the event and the requirement or purpose for the process step. Also included are the data and time (or range of data and time) on or over which the process step occurred, along with an identification of, and means of communication with, person(s) and organization(s) associated with the process step.
<b>Quantitative attribute accuracy</b>	Accuracy of quantitative attributes.	Metrics may involve the same procedures as those for positional accuracy. Must distinguish between single-valued quantitative thematic attributes and multiple-valued quantitative thematic attributes. The single-valued quantitative attribute is a measure to indicate how far the assigned value is from the best estimation value (the mean of all observations). The accuracy metrics for this may be expressed by root mean square error or standard deviation, bias, systematic error, range (min and max error), histogram of deviation from the mean value or by confidence interval and confidence level.
		The multiple-valued quantitative thematic attributes can be expressed by a list of accuracy values for single valued quantitative thematic attributes, the correlation matrix or the eigen values of the correlation matrix, the correlation function or by the range (min and max error).

Name	Definition	Further details / Measured by
<b>Relative or internal accuracy</b>	Closeness of the relative positions of features in the scope to their respective relative positions accepted as or being true.	The accuracy of positions of occurrences of entities relative to each other (in both the horizontal (x,y) and vertical (z) dimensions), and the results of testing procedures.
<b>Resolution</b>	Level of detail expressed as a scale factor or a ground distance. [Included in the Identification Information class of ISO/TC 211 rather than in the Data Quality class.]	
<b>Result information</b>	Generalization of more specific result classes.	This class includes a Conformance Result which contains information about the outcome of evaluating the obtained value (or set of values) against a specified acceptable conformance quality level. Also included is a Quantitative Result which provides information about the value (or set of values) obtained from applying a data quality measure.
<b>Scope</b>	The specific data to which the data quality information applies.	The scope subclasses include the hierarchical level of the data, the information about the spatial, vertical and temporal extent of the data, and a detailed description about the level of the data specified by the scope.
<b>Semantic accuracy</b>	The quality with which geographical objects are described in accordance with the selected model.	Semantic accuracy refers to the pertinence of the meaning of the geographical object rather than to the geometrical representation. One aspect of semantic accuracy could be termed "textural fidelity". One measure of textural fidelity could indicate the accuracy of spelling, perhaps by the percentage of wrong spellings. Another measure could be of the use of exonym or alternative spellings. A third metric may indicate the consistency of abbreviations.
<b>Source information</b>	Information about the source data used in creating the data specified by the scope.	Lists the name and the organization responsible for the data set, as well as the purpose, date, method of data capture, source documents, if appropriate, and creator of the data set's original production, including their address.

Name	Definition	Further details / Measured by
<b>Temporal accuracy</b>	Accuracy of the temporal attributes and temporal relationships of features.	Describes the correctness of time and updating of a data (sub)set (currentness) by such metrics as the moment of last update (in case of creation, modification, deletion or unchanged use), rate of change of entities per unit of time, trigger value (indicating the number of changes before a new version of the data set is issued), temporal lapse (giving the average time period between the change in the real world and the updating of the database) or temporal validity (indicating data to be out of date, valid or not yet valid).
<b>Temporal consistency</b>	Correctness of ordered events or sequences, if reported.	
<b>Temporal validity</b>	Validity of data specified by the scope with respect to time.	
<b>Thematic accuracy</b>	Accuracy of quantitative attributes and the correctness of non-quantitative attributes and the classification of features and their relationships.	Accuracy of continuous (i.e. scalar or quantitative) and discrete (i.e. qualitative or nominal) values associated with a feature or relationship in the data set. Accuracy values are given in the same unit and reference system as in the measured values.
<b>Topological consistency</b>	Correctness of the explicitly encoded topological characteristics of the dataset as described by the scope.	A metric for topological consistency in one dimension could be reported by indicating the percentages of junctions that are not formed when they should be, or in two dimensions by indicating the percentage of incorrectly formed polygons.
<b>Usage</b>	Any previous use of a data set by other users. [Included in the Identification Information class of ISO/TC 211 rather than in the Data Quality class.]	For each usage a separate statement should be given indicating the organization that has used the data set, the type of usage and its perceived fitness and any possible constraints or limitations that were imposed or discovered during that use.

### 9.1.A comparison of spatial dimensions

	<b>SDTS</b> (Spatial Data Transfer Standard)	<b>ICA</b> (International Cartographic Association)	<b>CEN/TC287</b> (European Committee for Standardization)	<b>FGDC</b> (Federal Geographic Data Committee)	<b>ISO/TC 211 (19115)</b> (International Standards Organization)
<b>Data Quality Information</b>	Lineage	Source	Source	Lineage (Source Information, Process Step)	Lineage, Process Step Information, Result Information, Scope, Source Information
			(Potential) usage		
<b>Data Quality Element Information</b>	Resolution	Resolution			Resolution
	Positional Accuracy	Metric accuracy	Metric accuracy	Positional Accuracy	Positional Accuracy
	Attribute Accuracy	Thematic accuracy	Thematic accuracy	Attribute Accuracy	Thematic accuracy
	Completeness	Completeness	Completeness	Completeness	Completeness
	Logical Consistency	Logical consistency	Logical consistency	Logical consistency	Logical consistency
		Semantic accuracy	Meta quality		
		Temporal accuracy	Temporal accuracy		Temporal accuracy
			Homogeneity		
			Cloud cover		

<b>Spatial Data Standards</b>
CEN/TC 287 (1997) As reported in Aalders (2002) & Jakobsson (2002) [Note: CEN is the European Committee for Standardization]
FGDC (2002) "Content Standard for Digital Geospatial Metadata Workbook. Version 2.0." Federal Geographic Data Committee.
ICA (1996) As reported in Aalders (2002) [Note: ICA = International Cartographic Association]
ISO/TC 211 [19155] (2001) "Geographic information / Geomatics - Metadata" . International Standards Organization. ISO/TC 211 Secretariat. N 1142.
SDTS (1992) Spatial Data Transfer Standard. ANSI NCITS 320-1998. ( <a href="http://mcmcweb.er.usgs.gov/sdts/standard.html">http://mcmcweb.er.usgs.gov/sdts/standard.html</a> )

## 9.2. MoFR spatial data quality dimensions details

The following spatial dimensions have been identified as having a high priority within MoFR. Details for each of the dimensions follows.

Dimension	Measurement Type	Keywords
Positional Accuracy	Automated if digitally available	Accuracy, metrics of error
Thematic Accuracy	Automated if digitally available	Quantitative Data, Qualitative Data, classification errors
Completeness	Generally manual but can check for related records	Measured in space, time or theme
Logical Consistency	Automate within GIS	Topological errors
Temporal Accuracy	Generally manual	Temporal attributes and relationship of features, metrics, rate of change, validity

### 9.2.1. Positional Accuracy

#### ○ Definition

Accuracy of the position of features.

#### ○ Measure By

- For points, accuracy is defined in terms of the distance between the encoded location and "actual" location.
- Error can be defined in various dimensions: x, y, z, horizontal, vertical, total.
- Metrics of error are extensions of classical statistical measures (mean error, RMSE or root mean squared error, inference tests, confidence limits, etc.) .
- For lines and areas, the situation is more complex. This is because error is a mixture of positional error (error in locating well-defined points along the line) and generalization error (error in the points selected to represent the line). The epsilon band is usually used to define a zone of uncertainty around the encoded line, within which "actual" line exists with some probability.

#### ○ Automation

Can be automate if reference system is available

### 9.2.2. Thematic Accuracy

#### ○ Definition

Accuracy of quantitative attributes and the correctness of non-quantitative attributes and the classification of features and their relationships.

#### ○ Measure By

- The metrics used here depend on the measurement scale of the data.
  - (1) Quantitative data (e.g., precipitation) can be treated like a z-coordinate (elevation) and assessed using metrics normally used for vertical error (such as the RMSE).
  - (2) Qualitative data (e.g., land use/land cover) is normally assessed using a cross-tabulation of encoded and "actual" classes at sample of locations. This produces a classification error matrix.

- **Automation**

Can be automate if reference system is available

### **9.2.3. Completeness**

- **Definition**

Presence and absence of features, their attributes and their relationships.

- **Measure By**

- It is assessed relative to the database specification, which defines the desired degree of generalization and abstraction (selective omission).
- Completeness can be measured in space, time, or theme.

- **Automation**

Generally manual

### **9.2.4. Logical Consistency**

- **Definition**

Degree of adherence to logical rules of data structure, attribution and relationships (data structure can be conceptual, logical or physical)

- **Measure By**

- Consistency is a measure of the internal validity of a database, and is assessed using information that is contained within the database.
- For line or polygon data, check for these types of errors: Dangles/unclosed polygons, Slivers, Duplicate lines, Intersections without nodes, Polygons with more than one label, and Adjacent polygons with the same attribute.

- **Automation**

Automated within GIS or customized reconciliation tool

### **9.2.5. Temporal Accuracy**

- **Definition**

Accuracy of the temporal attributes and temporal relationships of features.

- **Measure By**

- Metrics as the moment of last update (in case of creation, modification, deletion or unchanged use), rate of change of entities per unit of time, trigger value (indicating the number of changes before a new version of the data set is issued), temporal lapse (giving the average time period between the change in the real world and the updating of the database) or temporal validity (indicating data to be out of date, valid or not yet valid).

- **Automation**

Generally manual

## 10. Appendix D – Details of Business Rules

### The Business Rules of Data Quality

Data quality is the measurement of data against independently-defined business rules that describe what good quality data is supposed to look like. These rules are collected from a variety of sources around the organization, and once collected, they are then recorded and maintained in the central metadata repository. Some common sources to consult for defining business rules are:

- Subject matter experts
- User reference guides
- System documentation
- Data dictionaries
- Metadata repositories
- Data entry instructions
- Program code
- Intuition and experience

Business rules can be used to describe, for each column in the database, when and how the column is to be populated and how each piece of data is related to other pieces of data. The development of business rules against which data can be validated is typically an iterative process:

- a. Define a rule;
- b. Test the rule against the data;
- c. Analyze the test results;
- d. Refine the rule, as required;
- e. Re-test the refined rule.

The business rules are of the following general types:

#### VALUE FORMAT OR STRUCTURE

Some data elements have business rules that specify the structure of the data. A data element containing Social Insurance Numbers is an example of a structure business rule. Others might include telephone numbers or dates stored as character strings. Other business rules may be stated as exclusionary; such as person names should not contain certain special characters.

#### VALID VALUE SETS OR RANGES

The allowed value sets of some columns are bounded by a finite list of valid values or acceptable range of values. The domain of the valid Canadian postal codes is an excellent example of a valid value set business rule. Or a range of values between 1 and 365 bound a column that contains the day of the year, unless it's a leap year of course.

#### PRIMARY KEY UNIQUENESS

A primary key is the identifier of a set of data. An account number is the identifier for all of the information that pertains to that account and allows that information to be located in the table of information that describes all accounts. An employee ID identifies an individual working for the company.

In some instances, the primary key is comprised of two or more columns that are taken together as the identifier of the set of data. A table of the individual line items of an order might use the order number and the line item number taken together to be the primary key of the set of data describing the item of the order.

Regardless of whether the key is made up of one column or multiple columns, the primary key business rule says that each primary key value should appear once and only once in each table. As a real life example of the reason, you might be more than a little miffed if, when you paid off your home, the bank applied the payment against another mortgage with the same loan number.

It should be noted that most of the MoFR databases have defined primary keys and Oracle enforces the uniqueness of these primary keys and this testing may not be required in those databases. However, data being sent in from a lumber company may not be resident in a database and the primary key uniqueness of the data should be tested.

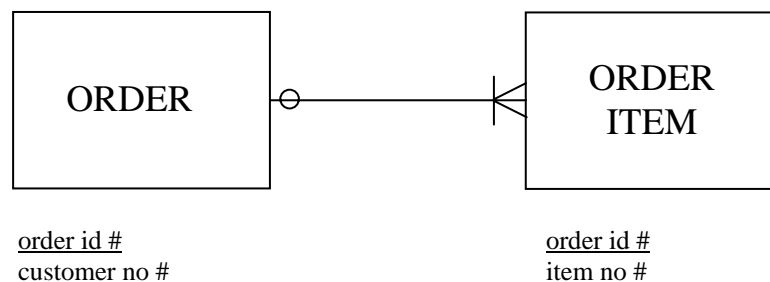
## ROW UNIQUENESS

Just as it is an error if the same loan number identifies two different loans, the reverse is also a problem. Consider what would happen if your loan was recorded twice with two different loan numbers. You could be paid up to date under one number and be getting foreclosure notices under the other. Since there can only be one set of data that describes each person, tract of land, event or whatever, that set of data should appear once and only once within the domain of the table values.

The data quality ideal to target is a one-to-one relationship between a primary key and the set of data that the primary key identifies. Each primary key is unique and each identifies a unique set of descriptive data. The row uniqueness identifiers are sometimes referred to as the 'natural key' of a table.

## REFERENTIAL INTEGRITY

Having primary keys identify sets of data makes it possible to support relationships between those sets of data.



It isn't necessary for each individual line item of an order to record the customer number, because the order number that, with item number is the primary key of the order item table is also the primary key of the order table where the customer number is stored.

The order number in the order item table is a foreign key of the order table's primary key. When a relationship is established, the primary key of the parent cascades to the child table as a foreign key.

The business rule of referential integrity simply states that all of the foreign key values of the child table should appear in the value set of the primary key of the parent table. If all values appear, the two tables are said to have referential integrity.

It should be noted that most of the MoFR databases have defined referential integrity constraints and Oracle enforces the uniqueness of these primary keys and this testing may not be required in those databases. Once again though, data being sent in from a lumber company may not be resident in a database and the referential integrity of the data should be tested.

These last three classes of business rules, the ones that deal with uniqueness of primary keys and row uniqueness identifiers and foreign key usage are referred to as structural rules. As a group, they allow us to assess the structural integrity of the tables of a database. These types of errors tend to be fairly rare, but are of extreme importance if they are found as they indicate a breakdown in the structural rules of the data.

## DEPENDENCY CONSTRAINTS

There are many instances where the values in one column are dependent upon the values in some other column or columns in the same table or in other related tables. There is almost an unlimited variety to these types of dependency rules, but most tend to fall into four basic classes:

### SIMPLE MATHEMATICAL

- $A (+ - * / ) B$  should equal  $C$  in the same row of the table
- Hours Worked times Pay Rate should equal Gross Pay
- The sum of On Order plus Shipped should equal Original Order Quantity
- The Principal Balance Due should equal Original Loan Amount minus Accumulated Principal Payments

### TWO TABLE MATHEMATICAL

- For each value of  $K$  in one table,  $A (+ - * / ) B$  should equal  $C$  in another table when matched on the value of  $K$
- The total of  $A$  for each subset value of  $K$  in one table should equal  $C$  in another table when matched on the value of  $K$
- The sum of the Order Value for each Order ID in the Order Item table should equal the Total Order Value in the Order Header table for that Order ID

### FUNCTION APPLICATIONS

- $A$  should equal function ( $B$ ) where the function is either built-in or rule defined by the business
- UpName should equal Upper(Last Name)

- CustAbbrev should equal the first four positions of Customer Name
- Sales Group should equal Region concatenated to District

#### IF... THEN... ELSE...

- If the value in *A* is *x* then *y* should be true, else *z* should be true
- If Past Due Days > 30 then New Charge OK Flag should be 'N' else 'Y'
- If the Foreign Supplier Flag is 'Y' then there should be a row in the Customs Expediter table for this Order ID

#### VALIDITY TABLE

- Valid value combinations from two or more columns that are used to test the validity of source data
- While Region '12' is valid and District '63' is valid and Salesman '425' is valid, the combination of 12-63-425 on an order is not

#### CARDINALITY

- Restrictions placed on the number of times that a foreign key value can appear in the value set off a child table.
- Each order must have at least one order item.
- There is a maximum of ninety-nine items per order.

#### DATA POPULATION

Data population rules tell us whether a column can contain null values, or blanks or zeroes depending upon the data type, or some other default value to represent a “don't know” condition. It was common legacy system design to use strings of 9's to represent logical states of the data such as last record, deletion flagged, or a non-expired expiration date.

All of these conditions can represent absences or voids of useful information having been recorded. Measuring the voids gives us an indication of the completeness of the data in a column.

#### RELATED VALUES

Complex business rules specific to a business area may define related values. For example, if harvesting is complete for a cut block, an opening must exist.

#### DATA TRANSFORMATIONS

One frequently occurring task is the movement of data from one environment to another as may be encountered in the population of ERP databases or the loading of a data warehouse. In some installations, this transformation can be a part of the nightly schedule as data from one legacy environment is moved to another dissimilar environment or into an in-progress ERP implementation.

In order to facilitate this movement of data from one environment to another, a class of commercial software specifically tuned to optimize this process has come into existence. The class is generally referred to as Extraction, Transformation and Loading software or ETL. Some of the more prominent names in this market are Informatica, Oracle, IBM Web Sphere, and Ab Inicio.

The data transformation business rule basically asks to prove that the data that existed in a source environment has been moved to the target environment according to the transformation rules. To test the validity of data transformation business rules, it is necessary to answer three questions:

- Are any target rows missing?
- Do any target rows have an unidentifiable source?
- Has the data in each row in each column been transformed according to the rules?

## ACCURACY

While the above classes of business rules go a long way towards defining quality data, they all share one common shortcoming. Each of these rules defines a means of testing data in terms of its structure or content. But what none of them talk about is the accuracy of the data; whether it is a correct representation of the real world facts about this person, thing, event or whatever. For example, in the absence of other data to validate it, it is difficult to detect that the birth year value that was recorded as 1976 should really have been 1975.

About the only possible way of testing for these types of errors is to compare the recorded data to the real world as found in the original document or other source, or through customer or employee “help us validate your account” surveys. All of these methods are expensive to conduct for a limited impact on the database.

By instilling the notion that data quality is important to MoFR and by taking the steps to ensure the entry, maintenance and usage of complete and structurally correct corporate data, accuracy will follow.

## 11. Appendix E – Data Quality Methodologies

Nine Data (Information) Quality Methodologies were examined in detail. A brief description of each methodology is given below and a comparison of all the methodologies to the proposed MoFR methodology is given in the table which follows.

- (1) **Six Sigma** is a disciplined, data-driven approach and methodology for eliminating defects (driving towards six standard deviations between the mean and the nearest specification limit) in any process -- from manufacturing to transactional and from product to service. (**DMAIC** = Design, Measure, Analyze, Improve, Control)
- (2) **Total Information Quality Management (TQM)**. A full methodology for information quality management as outlined by Larry English.
- (3) **AIM quality (AIMQ)** is a methodology that forms a basis for IQ assessment and benchmarking. (AIMQ = Acceleration, Integration Management Quality)
- (4) Among the most widely used tools for continuous improvement is a four-step quality model—the plan-do-check-act (**PDCA**) cycle, also known as Deming Cycle or Shewhart Cycle: According to Masaaki Imai, founder of the Kaizen Institute of Europe, "Kaizen simply means continuous improvement involving everybody in the organization. I think the two key words, one is improvement and the other is continuous."
- (5) **Quality Function Deployment (QFD)** is a systematic process for motivating a business to focus on its customers. It is used by cross-functional teams to identify and resolve issues involved in providing products, processes, services and strategies which will more than satisfy their customers.
- (6) The **House of Quality** is the first matrix in a four-phase QFD (Quality Function Deployment) process. It's called the House of Quality because of the correlation matrix that is roof shaped and sits on top of the main body of the matrix. The correlation matrix evaluates how the defined product specifications optimize or sub-optimize each other.
- (7) **Total Quality Management (TQM)** is a management approach that originated in the 1950's and has steadily become more popular since the early 1980's. Total Quality is a description of the culture, attitude and organization of a company that strives to provide customers with products and services that satisfy their needs. The culture requires quality in all aspects of the company's operations, with processes being done right the first time and defects and waste eradicated from operations. A core concept in implementing TQM is Deming's 14 points.
- (8) **Data Quality Practice**. Provides the entire lifecycle of data quality – what “quality” data are, who owns it and has responsibility for it, how to assess your data and measure ongoing quality, root cause analysis, metadata among other aspects of data quality management.
- (9) **ISO 9001:2000** specifies requirements for a quality management system for any organization that needs to demonstrate its ability to consistently provide product that meets customer and applicable regulatory requirements and aims to enhance customer satisfaction.

MoFR	Six Sigma DMAIC Phases & Steps [7, 5]	TIQM Processes & Steps [7,5]	AIMQ Methodology [11]	PDCA / Kaizen [13, 16, 5]	QFD [4,3,2]	House of Quality [2,3,5]	TQM [9,3,5]	Data Quality Practice [12]	ISO 9000:2000 [14]
<b>Identify Problems</b>	<ul style="list-style-type: none"> <li>○ Identify the problem</li> <li>○ Define requirements</li> <li>○ Analyze priorities</li> </ul>	<ul style="list-style-type: none"> <li>○ Define the project</li> <li>○ Plan the objectives</li> <li>○ Identify impact</li> </ul>	<ul style="list-style-type: none"> <li>○ Develop questionnaire</li> </ul>	<ul style="list-style-type: none"> <li>○ Plan</li> <li>○ Identify the problem</li> <li>○ Set a goal</li> <li>○ Establish governance</li> </ul>	<ul style="list-style-type: none"> <li>○ Design</li> </ul>	<ul style="list-style-type: none"> <li>○ Design</li> <li>1. Customer Requirements ("Voice of the Customer")</li> <li>2. Regulatory Requirements: Document requirements dictated by management and/or regulatory standards</li> <li>3. Customer Importance Rating. Rate importance of each requirement.</li> <li>4. Customer Rating of Competition</li> </ul>	<ul style="list-style-type: none"> <li>○ Select the theme</li> <li>○ Define the problem</li> </ul>	<ul style="list-style-type: none"> <li>○ Identify the problem</li> <li>○ Define requirements</li> </ul>	<ul style="list-style-type: none"> <li>○ Define requirements</li> <li>○ Establish quality policy</li> </ul>
	<ul style="list-style-type: none"> <li>○ Plan</li> <li>○ Measure performance</li> <li>○ Develop baseline</li> </ul>	<ul style="list-style-type: none"> <li>○ Assess Data &amp; Information Quality</li> <li>○ Assess costs</li> </ul>	<ul style="list-style-type: none"> <li>○ Assess conformance to specifications</li> <li>○ Do gap analysis</li> </ul>	<ul style="list-style-type: none"> <li>○ Plan</li> <li>○ Identify and analyze processes</li> </ul>	<ul style="list-style-type: none"> <li>○ Analyze processes and performance</li> </ul>	<ul style="list-style-type: none"> <li>5. Technical Descriptors ("Voice of the Engineer"). Obtain attributes that can be measured and benchmarked.</li> <li>6. Direction of Improvement. Direction of movement for each descriptor.</li> <li>7. Relationship matrix. Relationship between customer need &amp; ability to meet that need.</li> <li>8. Technical Analysis of Competition. Examine technical descriptors of competition</li> <li>9. Target Values for Technical</li> </ul>		<ul style="list-style-type: none"> <li>○ Map the Information Chain</li> <li>○ Establish data quality scorecard</li> <li>○ Assess costs</li> </ul>	<ul style="list-style-type: none"> <li>○ Plan objectives</li> <li>○ Plan &amp; develop realization processes</li> <li>○ Document system</li> <li>○ Control development</li> </ul>

MoFR	Six Sigma DMAIC Phases & Steps [7, 5]	TIQM Processes & Steps [7,5]	AIMQ Methodology [11]	PDCA / Kaizen [13, 16, 5]	QFD [4,3,2]	House of Quality [2,3,5]	TQM [9,3,5]	Data Quality Practice [12]	ISO 9000:2000 [14]
						Descriptors. 10. Correlation matrix (the "roof" of the house). How descriptors impact each other. 11. Calculate importance. Calculate absolute importance for each technical descriptor.			
<b>Analyze Solutions</b>	<ul style="list-style-type: none"> <li>○ Identify root causes</li> </ul>	<ul style="list-style-type: none"> <li>○ Identify root causes</li> </ul>	<ul style="list-style-type: none"> <li>○ Analyze gap analysis</li> </ul>	<ul style="list-style-type: none"> <li>○ Plan</li> <li>○ Identify root causes</li> </ul>	<ul style="list-style-type: none"> <li>○ Process planning matrix</li> <li>○ Concept selection matrix</li> </ul>	.	<ul style="list-style-type: none"> <li>○ Analyze the problem</li> </ul>	<ul style="list-style-type: none"> <li>○ Identify root causes</li> <li>○ Build project team</li> <li>○ Build vs buy</li> <li>○ Define DQ rules</li> </ul>	<ul style="list-style-type: none"> <li>○ Identify personnel</li> <li>○ Identify infrastructure</li> <li>○ Provide quality environment</li> <li>○ Control purchasing</li> </ul>
<b>Implement changes</b>	<ul style="list-style-type: none"> <li>○ Implement solutions</li> <li>○ Test solutions</li> <li>○ Standardize solutions</li> </ul>	<ul style="list-style-type: none"> <li>○ Plan improvement</li> <li>○ Implement improvements</li> <li>○ Check impact</li> <li>○ Measure costs</li> <li>○ Standardize improvements</li> </ul>	<ul style="list-style-type: none"> <li>○ Test solutions with gap analysis</li> </ul>	<ul style="list-style-type: none"> <li>○ Do</li> <li>○ Develop solution</li> <li>○ Implement solution</li> <li>○ Check</li> <li>○ Check solution</li> <li>○ Standardize solution</li> </ul>	<ul style="list-style-type: none"> <li>○ Production</li> </ul>		<ul style="list-style-type: none"> <li>○ Generate ideas</li> <li>○ Test ideas</li> <li>○ Implement ideas</li> <li>○ Check results</li> <li>○ Standardize</li> </ul>	<ul style="list-style-type: none"> <li>○ Execute improvement</li> </ul>	<ul style="list-style-type: none"> <li>○ Perform remedial processes</li> <li>○ Establish quality system</li> <li>○ Develop management system</li> <li>○ Monitor and measure</li> </ul>
<b>Monitor Results</b>	<ul style="list-style-type: none"> <li>○ Establish standard to maintain performance</li> <li>○ Correct problems as needed</li> <li>○ Celebrate success</li> </ul>	<ul style="list-style-type: none"> <li>○ Standardize quality improvements</li> <li>○ Report on improvements</li> <li>○ Improve process</li> <li>○ Document</li> </ul>	<ul style="list-style-type: none"> <li>○ Measure against others or best practices with gap analysis</li> <li>○ Use results to continually improve</li> </ul>	<ul style="list-style-type: none"> <li>○ Act</li> <li>○ Ongoing monitoring</li> <li>○ Refine solutions</li> <li>○ Look for other opportunities</li> </ul>				<ul style="list-style-type: none"> <li>○ Measure improvement</li> <li>○ Build on success</li> </ul>	<ul style="list-style-type: none"> <li>○ Analyze quality information</li> <li>○ Improve quality management system</li> <li>○ Prevent potential nonconformities</li> </ul>

