Pitney Bowes
Business Insight

Five Fundamental Data Quality Practices

WHITE PAPER:

DATA QUALITY & DATA INTEGRATION

David Loshin

INTRODUCTION

DATA QUALITY MANAGEMENT INCORPORATES A "VIRTUOUS CYCLE" IN WHICH CONTINUOUS ANALYSIS, OBSERVATION, AND IMPROVEMENT LEAD TO OVERALL IMPROVEMENT IN THE QUALITY OF ORGANIZATIONAL INFORMATION ACROSS THE BOARD. THE RESULTS OF EACH ITERATION CAN IMPROVE THE VALUE OF AN ORGANIZATION'S DATA ASSET AND THE WAYS THAT DATA ASSET SUPPORTS THE ACHIEVEMENT OF BUSINESS OBJECTIVES.

THIS CYCLE TURNS ON THE EXECUTION OF FIVE FUNDAMENTAL DATA QUALITY MANAGEMENT PRACTICES, WHICH ARE ULTIMATELY IMPLEMENTED USING A COMBINATION OF CORE DATA SERVICES. THOSE PRACTICES ARE:

- DATA QUALITY ASSESSMENT
- DATA QUALITY MEASUREMENT
- INTEGRATING DATA QUALITY INTO THE APPLICATION INFRASTRUCTURE
- OPERATIONAL DATA QUALITY IMPROVEMENT
- DATA QUALITY INCIDENT MANAGEMENT

BY ENABLING REPEATABLE PROCESSES FOR MANAGING THE OBSERVANCE OF DATA QUALITY EXPECTATIONS, THESE PRACTICES PROVIDE A SOLID FOUNDATION FOR ENTERPRISE DATA QUALITY MANAGEMENT. THIS PAPER DESCRIBES THESE PRACTICES AND THEN LOOKS AT THE CORE DATA SERVICES UPON WHICH THESE PRACTICES RELY. EACH SECTION WILL PROVIDE AN OVERVIEW OF THE PRACTICE AND REVIEW PROCESSES THAT ARE USED TO ACHIEVE THE DESIRED PRACTICE'S OBJECTIVES. BY COMBINING GOOD DATA MANAGEMENT PRACTICES WITH THE RIGHT TECHNOLOGY PLATFORM, AN ORGANIZATION CAN FULLY INCORPORATE DATA QUALITY INTO THE ENTERPRISE ARCHITECTURE.

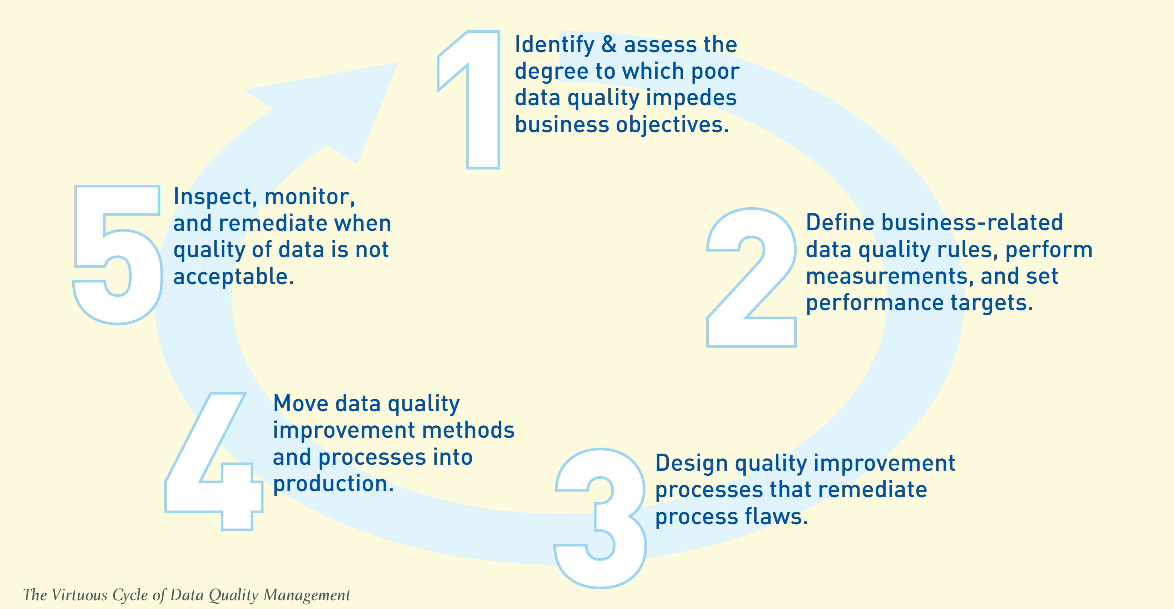
DATA QUALITY MANAGEMENT PRACTICES CAN IMPROVE THE VALUE OF AN ORGANIZATION'S DATA ASSET AND THE WAYS IT SUPPORTS THE ACHIEVEMENT OF BUSINESS OBJECTIVES

The Virtuous Cycle of Data Quality

Data quality management incorporates a "virtuous cycle" in which continuous analysis, observation, and improvement lead to overall improvement in the quality of organizational information across the board (see below). The objective of this cycle is to transition from being an organization in which the data stewards react to acute data failures into an organization that proactively controls and limits the introduction of data flaws into the environment.

In turn, this virtuous cycle incorporates five fundamental data quality management practices, which are ultimately implemented using a combination of core data services. Those practices are:

1. Data quality assessment, as a way for the practitioner to understand the scope of how poor data quality affects the ways that the business processes are intended to run, and to develop a business case for data quality management;
2. Data quality measurement, in which the data quality analysts synthesize the results assessment and concentrate on the data elements that are deemed critical based on the selected business users' needs. This leads to the definition of performance metrics that feed management reporting via data quality scorecards;
3. Integrating data quality into the application infrastructure, by way of integrating data requirements analysis across the organization and by engineering data quality into the system development life cycle;
4. Operational data quality improvement, where data stewardship procedures are used to manage identified data quality rules, conformance to acceptability thresholds, supported by
5. Data quality incident management, which allows the data quality analysts to review the degree to which the data does or does not meet the levels of acceptability, report, log, and track issues, and document the processes for remediation and improvement.


1
Identify & assess the degree to which poor data quality impedes business objectives.
2
Define business-related data quality rules, perform measurements, and set performance targets.
3
Design quality improvement processes that remediate process flaws.
4
Move data quality improvement methods and processes into production.
5
Inspect, monitor, and remediate when quality of data is not acceptable.

The Virtuous Cycle of Data Quality Management

Together, these practices establish the foundation of the data quality management program, since they enable a repeatable process for incrementally accumulating metrics for data quality that will contribute to populating a data quality scorecard, a data quality dashboard, as well as driving proactive data quality management. In turn, trained staff must employ core data services to make these practices operational.

Data Quality Assessment

Smart organizations want to maximize their investment in data quality management, and this means understanding how poor data quality negatively impacts the achievement of business objectives. By quantifying that value gap, the data quality practitioner can determine the cost-effectiveness, feasibility, and speed of any proposed data quality improvement. Understanding the impacts of data flaws within the context of the business helps provides a yardstick to measure and prioritize emergent data issues.

As an example, there may be some suspicion of increased mailing and shipping costs due to inaccurate or invalid addresses. This suspicion may be introduced by a perception of a large number of undelivered shipped items returned. However, invalid or incorrect addresses not only incurs direct costs associated with returned items; analytical applications used to profile customer purchase patterns by region are skewed, which can impact the effective execution of marketing campaigns and regional sales promotions. The data quality assessment process can be used to quantify those costs and impacts and determine what percentage of those costs is directly attributed to addresses that can be corrected.

This practice incorporates processes for identifying, assessing, quantifying, and prioritizing data quality issues:

- Business Impact Analysis – This process is intended to guide the analysts by noting any potential data-related issues that increase costs, reduce revenues, impact margins, or introduce inefficiencies or delays in business activities. In essence, the objective is to identify any negative business impacts that can be attributed to data of unacceptable quality. Identifying the location and magnitude of critical paint points in the various business processes helps to scope the business requirements for information for the assessment, narrow the list of data sets that will be examined, and guide the identification of data quality requirements.
- Data Quality Assessment using Data Profiling – This process performs a bottom-up review of the actual data as a way to isolate apparent anomalies that may be real data flaws. Using data profiling and other statistical and analysis techniques, the analysts can identify these apparent anomalies, which can be subjected to further scrutiny when reviewed with business data consumers.
- Data Quality Assessment Anomaly Review – During this process, the data quality analysts review the discovered apparent anomalies with business data consumers to see if there are any links between the data errors and any potential business impacts. By distinguishing those data errors that have material impact from the irrelevant ones, the team can prioritize issues based on business impact, and explore ways that the issues can be resolved.
- Define Measures of Data Quality – Correlating business impacts to data issues through defined business rules provides the method of measurement, and these measures can be used to baseline levels of data quality as well as continuous observation and inspection within an information production flow. This process guides the consideration of data measures to be performed and the technology requirements for collecting those measurements.

THE OBJECTIVE IS TO ACHIEVE PROACTIVE CONTROLS AND THEREFORE LIMIT THE INTRODUCTION OF DATA FLAWS INTO THE ENVIRONMENT

- Prepare DQ Assessment Report – The process of documenting the correlation of business impacts with data anomalies along with potential methods of measurement all within a single report provides a "fix-point" for the business data consumers regarding the current state of data quality, and provides the baseline for considering target levels for improvement.

Data Quality Measurement and Metrics

Having used an assessment to identify areas for data quality improvement, the next step is to synthesize the results of the assessment to narrow the scope by concentrating on the data elements that are deemed critical based on the business users' needs. Defining performance metrics for reporting using a data quality scorecard requires processes for the determination of dimensions and corresponding units of measure and acceptability thresholds, and the presentation of quantifiable metrics that are relevant to the business data consumers.

To continue our example, once we have determined using the data quality assessment process that problems with addresses impacts the ability to optimally deliver shipped items, we can narrow the focus for data quality measurements to specific metrics associated with the critical data elements that contribute to the delivery failures. Some items might not be delivered due to missing street information, while others might have incorrect ZIP codes. The first problem is one of completeness, while the second of consistency with defined reference data. Measurements associated with the data quality dimensions of completeness and consistency can be defined using data quality validation rules for each address, and the resulting measures can be presented as metrics to the business users in the fulfillment department to estimate how invalid addresses are related to increased costs.

Aspects of this practice include:

- Select Dimensions of Data Quality – A dimension of data quality describes a context and a frame of reference for measurement along with suggested units of measurement. Commonly measured dimensions of data quality include completeness, consistency, timeliness, and uniqueness, although the range of possible dimensions is only limited by the ability to provide a method for measurement. During this process, the data quality analysts select the dimensions that are to be measured and consider the tools, techniques, and skills needed to capture the measurements. The result of this process is a collection of specific measures that can be combined to contribute to qualitative data quality metrics.

- Define Data Quality Metrics – Having identified the dimensions of data quality that are relevant to the business data consumers as well as the dimensions and the specific measures, the analyst can create specific reportable metrics that can be presented to the business data stewards. These may be basic metrics composed of directly measured rules, or may be more complex metrics that are composed as weighted averages of collected scores. Other aspects include reporting schemas and methods for drilling into flawed data for root cause analysis.

- Define Data Validity Rules – The assessment process will expose potential anomalies, which are reviewed with the business users to identify data quality measures and, ultimately, data quality metrics. Yet in order to transition away from a reactive approach that seeks to remediate data quality issues once they are manifested at the end-user interface, the organization must engineer data controls into the application development process so that data errors can be identified and addressed as they occur. This process has the data quality analysts developing data validity rules; these rules can be integrated into the business applications as controls to verify that data meet expectations throughout the information flow.

- Set Acceptability Thresholds – Once the data quality dimensions and metrics have been validated, the business users are consulted to express their acceptability thresholds. When a metric score is below the acceptability threshold, it means that the data does not meet business expectations. Integrating these thresholds with the methods for measurement completes the construction of the data quality metric.
- Devise Data Quality Scorecard – A data quality scorecard presents metric scores to the data stewards observing the business data sets. Metrics scores can be captured within a repository over a long time period to enable trending and demonstrate continuous improvement or (conversely) show that progress is not being made. The process of devising the scorecard include managing the metrics definitions, measurement processes, weightings, how the scores are captured and stored, as well as composing the tools and technologies for delivery and presentation.

Data Quality and the System Development Life Cycle

Too often, data quality becomes an afterthought, with staff members reacting to discovered errors instead of proactively rooting out the causes of data flaws. Because data quality cannot just be an afterthought, once there are processes for identifying the business impact of data quality as well as the capability to define rules for inspection and monitoring, the next step is to integrate that inspection directly into the business applications. In essence, the next practice is to establish the means by which data quality management is designed and engineered across the enterprise application architecture.

However, because traditional approaches to system requirements analysis and design have concentrated on functional requirements for transactional or operational applications, the information needs of downstream business processes are ignored until long after the applications are put into production. Instead, engineering data quality management into the enterprise requires reformulating the view to requirements analysis, with a new focus on horizontal and downstream information requirements instead of solely addressing immediate functional needs.

To continue our example, with the understanding that invalid addresses lead to increased shipping costs, there are two approaches for remediation. The reactive approach is to subject all addresses to a data cleansing and enhancement process prior to generating a shipping label as a way of ensuring the best addresses. While this may result in reducing some of the increased costs, there may be records that are not correctable, or are not properly corrected. Yet if the data validity rules are known, they can be integrated directly into the application when the location data is created. In other words, validating and correcting the address when it is entered by the customer prevents invalid addresses from being introduced into the environment altogether!

Processes that contribute to this practice include:

- Data Quality Requirements Analysis – During this process, the data quality analysts will synthesize data quality expectations for consumed data sets based on the business impact analysis, the determination of data quality dimensions, and aspects of prioritization related to feasibility as well as systemic impact. For each business application, the information flow is traversed backwards to the points where data is created or acquired, and the end-to-end map is investigated to determine the most appropriate points for inserting data inspection routines. At the points where data sets are extracted, transformed, or exchanged, the analysts can propose data controls that will trigger notification events when downstream expectations are violated.

STANDARDIZING THE WAY DATA QUALITY IS DEPLOYED AND USING THE RIGHT KINDS OF TOOLS WILL ENSURE PREDICTABLE INFORMATION RELIABILITY AND VALUE

- Enhancing the SDLC for DQ – Incorporating data validation and data quality inspection and reporting into business processes and the corresponding business application by adjusting the general system development life cycle (SDLC) so that organizational data requirements can be solicited and integrated into the requirements phase of system development. This process looks at business ownership of data, and how business process modeling can be used to elaborate on the information needs in addition to functional requirements for business operations. Since downstream users such as business intelligence reporting consumers will depend on the data collected during operational activities, there is a need to formally collect data requirements as part of the SDLC process.
- Integrate data quality improvement methods – Capturing the organization's data quality requirements as part of the requirements and design phases of a development life cycle empower the development team in integrating data quality and data correction directly into the application. This includes the ability to validate data values and records at their entry into the environment (either through acquisition or creation) or at any hand-off between processing stages, verify acceptability, and either push invalid data back to the provider for resolution or to apply adjustments or corrections on the fly.

Operational Data Quality Improvement

Having collected data quality requirements, defined data validation rules and recommended methods for measuring conformance, the next step is to establish the contract between data suppliers and data consumers as to the service level for maintaining high quality data.

In our example, addresses are validated against a set of defined data standards, either specifically managed by postal agencies in different countries, or "de facto" standards employed by delivery practitioners to ensure proper distribution. These data standards define reference tables and other metadata artifacts that can be used to actively ensure that the validity of a delivery location specification.

Combining the data validity rules and the documented metadata, the data quality analysts can document the level of acceptability for location data expected by the business users. In turn, the performance of any remediation activities can be measured to guarantee that the data is of acceptable quality.

The practice of establishing a data quality service level agreement incorporates these tasks:

- Data Standards Management – The absence of a common frame of reference, as well as common business term definitions and an agreed-to format for exchange makes it difficult for parties to understand each other. This is acutely true with respect to data when specific pieces of information need to be shared across two or more business applications. This suggests the need for a normalized standard for data sharing. A data standard is an agreement between collaborating parties on the definitions of common business terms, the ways those terms are named and represented in data, and a set of rules that may describe how data are stored, exchanged, formatted, or presented. This process describes the policies and procedures for defining rules and reaching agreement about standard data elements.
- Active Metadata Management – Because the use of the data elements and their underlying concepts drive how the business applications will ultimately execute, an enterprise metadata repository can be used as a "control center" for driving and managing how those business applications use common data concepts. Aside from the need to collect standard technical details regarding the numerous data elements that are potentially available, a metadata repository can help when there is a need to
  > determine business uses of each data element,

> determine which data element definitions refer to similar concepts,

> identify the applications that refer to those data concepts,

> review how each data element and associated concepts are created, read, modified, or retired by different applications,

> document the data quality characteristics, note the inspection and monitoring locations within the business process flow, and

> summarize how all the uses are tied together.

Therefore, a valuable component of an information architecture is an enterprise business metadata management program to facilitate the desired level of standards across the organization.

- Data Quality Inspection and Monitoring – The availability of data validation rules is the basis for data quality inspection and monitoring. Inserting inspection probes and monitoring the quality of data provides the means for identifying data flaws and for notifying the appropriate people when those data flaws are discovered so that any agreed-to remediation tasks can be initiated. Mechanisms for data inspection and monitoring and the corresponding process workflows must be defined for the purposes of inspecting data and ensuring that the data elements, records, and data sets meet downstream requirements.

This process involves defining the data quality inspection routines, which may include both automated and manual processes. Automated processes may include the results of edit checks executed during application processing, data profiling or data analysis automation, ETL tools, or customized processing. Manual inspection may require running queries or reports on data sources or even obtaining samples of data which are then examined. Inspection procedures are defined for each relevant data quality dimension. The inspection methods are customized for each system as appropriate.

- Define Data Quality Service Level Agreements – A service level agreement is a contract between a service provider and that provider's consumers that specifies the service provider's responsibilities with respect to different measurable aspects of what is being provided, such as availability, performance, response time for problems, etc. A data quality service level agreement, or DQ SLA, is an agreement that specifies data consumer expectations in terms of data validity rules and levels of acceptability, as well as reasonable expectations for response and remediation when data errors and corresponding process failures are discovered. DQ SLAs can be expressed for any situation in which a data supplier provides data to a data consumer.

This process is to specify expectations regarding measurable aspects relating to one or more dimensions of data quality (such as accuracy, completeness, consistency, timeliness, etc.), as suggested by other processes already described. Then, the service level agreement specifies what is meant by conformance to those expectations, and describes the workflow that is performed when those expectations are not met. Reported issues will be prioritized and the appropriate people in the organization will be notified to take specific actions to resolve issues before any negative business impacts can occur.

Issue Tracking, Remediation, Improvement

Operationalizing the data quality service level agreement means that there are processes for reporting, logging, and tracking emergent or discovered data quality issues. Incident reporting frameworks can be adapted to this purpose, which allows the data stewards to concentrate on evaluating the root causes of data issues and proposing a remediation plan, ranging from process reengineering to simple data corrections.

BY COMBINING GOOD DATA MANAGEMENT PRACTICES WITH THE RIGHT TECHNOLOGY PLATFORM, AN ORGANIZATION CAN FULLY INCORPORATE DATA QUALITY INTO THE ENTERPRISE ARCHITECTURE

Issues tracking, logging, and management ensures that any discovered issues don't fall through the cracks. In our example, any time a shipment is returned due to a data quality problem, a data analyst will review the error to determine the source of the problem, consider whether it was due to a validation step that was not taken, or determine that there is a new root cause that can lead to defining additional validation rules that can be integrated into the business process flow.

This practice incorporates these tasks:

- Data Quality Issue Reporting and Tracking – Enforcing a data quality service level agreement requires the processes for reporting and tracking data quality issues as well as any follow-on activities. Using a system to log and track data quality issues encourages more formal evaluation and initial diagnosis of "bad data," and the availability of a data quality issue tracking system helps staff members be more effective at identifying and consequently fixing data-related problems. Incident tracking can also feeds performance reporting such as mean-time-to-resolve issues, frequency of occurrence of issues, types of issues, sources of issues, and common approaches for correcting or eliminating problems.
- Root Cause Analysis – Data validation rules used as data controls integrated within business applications can trigger notifications that a data error has occurred. At that point it is the role of the data stewards to not just correct the data, but also identify the source of the introduction of the errors into the data. The root cause analysis process employs inspection and monitoring tools and techniques to help isolate the processing phase where the error actually occurred and to review the business processes to determine the ultimate root cause of any errors.
- Data Cleansing – Remedying data errors is instinctively reactive, incorporating processes to correct errors in order to meet acceptability limits, especially when the root cause cannot be determined or if it is beyond the administrative domain of the data stewards to influence a change to the process. Corrections must be socialized and synchronized with all data consumers and data suppliers, especially when the data is used in different business contexts. For example, there must be general agreement for changes when comparing reported data and rolled-up aggregate results to operational systems, because different numbers that have no explanation will lead to extra time spent attempting to reconcile the variant results.
- Process Remediation – Despite the existence of a governed mechanism for correcting bad data, the fact that errors occur implies that flawed processes must be reviewed and potentially corrected. Process correction encompasses governed process for evaluating the information production flow, business process work flow, and the determination of how processes can be improved so as to reduce or eliminate the introduction of errors.

Data Quality Practices and Core Data Services

Instituting a data quality management program means more than just purchasing data cleansing tools or starting a data governance board, and establishing a good data management program takes more than just documenting a collection of processes. An iterative cycle of assessment, planning, execution, and performance management for data quality requires repeatable processes that join people with the right sets of skills with the most appropriate tools, and the staff members who are to take part in the program need to have the right kinds of tools at their disposal in order to transition from theory to actual practice. This suggests a combination of the right technology and the proper training in the use of technology, employing data services such as:

- Data integration, to ensure suitable means for extracting and transforming data between different kinds of systems.

- Data profiling, used for data quality assessment, data validation, and inspection and monitoring.
- Parsing and standardization and identity resolution, which is used for data validation, identification of data errors, normalization, and data correction.
- Record Linkage and merging, also used to identify data errors and for resolving variance and subsequent data correction.

These are a subset of the core data services for standardizing sound data management practices. Standardizing the way data quality is deployed and using the right kinds of tools will ensure predictable information reliability and value. When developing or reengineering the enterprise architecture, implementing the fundamental data quality practices will ultimately reduce the complexity of the data management framework, thereby reducing effort, lowering risk, and leading to a high degree of trust in enterprise information.

PITNEY BOWES BUSINESS INSIGHT: YOUR SOURCE FOR ANSWERS AND SOLUTIONS

THE PITNEY BOWES SPECTRUM™ TECHNOLOGY PLATFORM				
Pitney Bowes Spectrum Enterprise Data Quality Solution	**Pitney Bowes Spectrum Enterprise Data Governance Solution**	**Pitney Bowes Spectrum Enterprise Location Intelligence Solution**	**Pitney Bowes Spectrum Enterprise Data Integration Solution**	**Pitney Bowes Spectrum Business Services**
Universal Addressing Module • Address Now Module • Data Normalization Module • Universal Name Module • Advanced Matching Module	Profiler Plus • Monitor Plus	Enterprise Geocoding Module • Location Intelligence Module • Enterprise Routing Module	Data Services for Oracle • Data Services for Siebel • Data Services for SFDC • Data Services for mySAP • Sagent Data Flow	Enterprise Tax Management • Enterprise Routing • Data Quality Connector for mySAP • Data Quality Connector for Siebel • Global Sentry

Pitney Bowes
Business Insight

UNITED STATES

One Global View
Troy, NY 12180
main: 518.285.6000
800.327.8627
www.pbinsight.com
pbbi.sales@pb.com

CANADA

26 Wellington Street East
Suite 500
Toronto, Ontario
M5E 1S2
1.800.268.DATA
www.pbinsight.ca
pbbi.canada.sales@pb.com

EUROPE/UNITED KINGDOM

Minton Place, Victoria Street
Windsor, Berkshire SL4 1EG
United Kingdon
+44.1753.848.200
www.pbinsight.co.uk
pbbi.europe@pb.com

ASIA PACIFIC/AUSTRALIA

Level 7, 1 Elizabeth Plaza
North Sydney NSW 2060
Australia
+61.2.9437.6255
pbbi.australia@pb.com
pbbi.sea@pb.com
pbbi.china@pb.com