

# Data Quality Considerations in Master Data Management Structures

WHITE PAPER:

**CUSTOMER DATA QUALITY PLATFORM**

Michael Overturf  
Navin Sharma



# Data Quality Considerations in Master Data Management Structures

2

## ABSTRACT

COMPANIES ACQUIRING COMPANIES. HUMAN RESOURCES SHARING INFORMATION WITH FINANCE. BUSINESSES SPANNING MULTIPLE COUNTRIES. WHAT DO ALL OF THESE SCENARIOS HAVE IN COMMON? THE SHARING OF DATA. WHAT IS THE CRITICAL NEED SHARED BY ALL OF THESE SCENARIOS? DATA MANAGEMENT. MASTER DATA MANAGEMENT, TO BE EXACT. IN TODAY'S BUSINESS ENVIRONMENT, MERGERS AND ACQUISITIONS, DATA SHARING, AND GLOBAL MARKET PRESENCE ARE COMMONPLACE. TO BE SUCCESSFUL, COMPANIES MUST NOT ONLY GAIN CONTROL OF, BUT MUST DRASTICALLY IMPROVE THEIR CUSTOMER DATA. TO DO THIS, IT IS PARAMOUNT FOR COMPANIES TO IMPLEMENT AND ENFORCE ENTERPRISE-WIDE MASTER DATA MANAGEMENT SOLUTIONS.

# CORRECT CUSTOMER INFORMATION IS THE LIFEBLOOD OF EVERY COMPANY

## What is Master Data Management?

Master Data Management (MDM) is a process for effectively creating a complete data context view in medium to large distributed data environments. MDM may be accompanied by Customer Data Integration (CDI) systems that provide a simplified access layer to corporate applications. The interplay between MDM and CDI provides system architects the facilities to ensure the operational contiguity of data. More simply put, it is the ability to understand the transactional context between a corporation and its customers.

However, while most business managers agree on the value of this knowledge, many lag in the adoption of requisite technologies to attain it. This paper discusses the continued intransigence of corporate data quality problems and how such data quality issues can be resolved within the MDM and CDI implementation context.

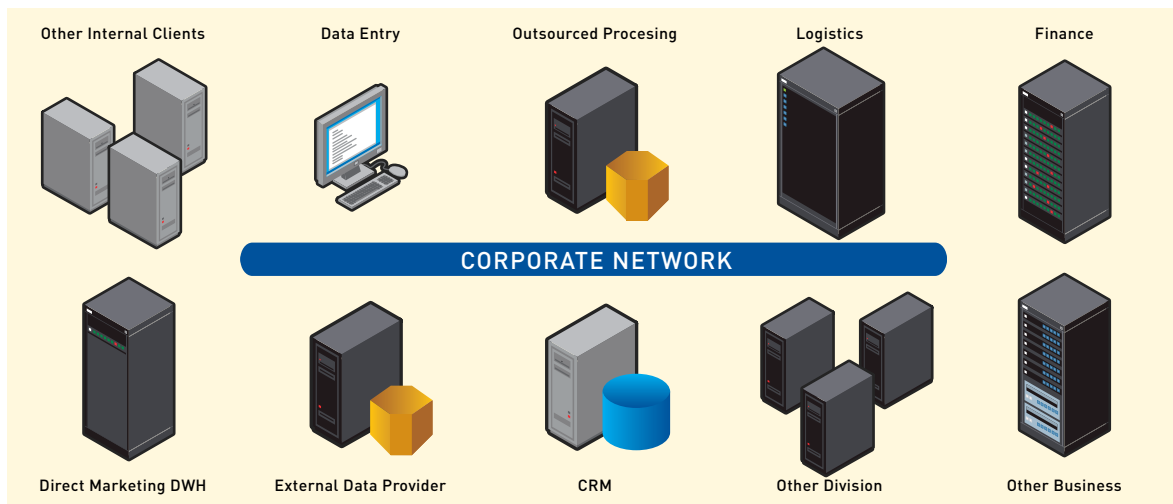
## State of the Art Data Quality Management

Most system architects know why data quality is important. Still, data management as a discipline is surprisingly underdeveloped in many companies. For example, in a

recent Ambysoft survey<sup>1</sup> of over 1,100 US IT organizations, only 40% of companies had some sort of developer regression test suite in place to validate data, yet over 95% affirmed their belief that data is a corporate asset. Two thirds (66%) of those respondents stated that their development groups would sometimes bypass their data management staff, if they had one, because they were 'too slow', 'too difficult', or 'offered too little value'.

Most companies are making inroads to improving their overall data quality. In a separate Ambysoft survey<sup>2</sup> of over 1,100 US companies, 52% stated that their data quality was 'pretty good, with a few problems', and 38% of those respondents called their data sources 'a complete mess'. But it is in the nitty-gritty that much remains to be done. Only 38% of respondents followed naming conventions for data, 58% either had no conventions or did not follow them consistently, 47% of respondents require more than a month to 'safely rename a production column', and 8% stated it was too risky to even try.

What can operational IT managers, data, and system architects do to improve quality against this background?



Typical IT environment that obviates the need for MDM

<sup>1</sup> Scott Ambler's August 2006 'Current State of Data Management' Survey  
<sup>2</sup> Scott Ambler's September 2006 Data Quality Survey

## Data Quality Considerations in Master Data Management Structures

4

### Architectural Challenges

Mid-to-large corporations have installed a host of specialized information systems that support various parts of the business. A corporate network could include a CRM system from Oracle, a transaction management system from IBM, a logistics management system from SAP, and a financial management system from Lawson – or any possible combination of the listed systems.

Each system has its own data definitions. For example, a CRM system may list customer name as [<First>, <Middle>, <Last>], while a legacy system will have an aggregated [<Name>] field, listed as [<Name1>, Name2>, <Name3>]. When attempting to combine this data into a comprehensive understanding of each customer, a programmer or database designer must arbitrate between many different interpretations and encodings of essentially the same, or similar, semantics. This can generate thousands of mappings and combinations. Without a manageable data quality rules database, this becomes a daunting situation.

Variation in update frequencies of each individual data system add further complexity. Online Transaction Processing (OLTP) systems continually update their data, but metadata updates are usually marked with undesirable service interruptions, as reflected in the above surveys. Cyclical systems, such as financial or sales management systems, update in different contexts from different sources.

Master database taxonomical relationships with their data source have three primary forms: a federated architecture, a data warehouse (DWH), and a hybrid. A federated architecture leaves the source data in its original location and uses a live, continuous access mechanism to read data, when requested. This type of architecture requires the correction of data errors in situ, at the source. A DWH copies data from the various sources and stores it for future

use, allowing the option to leave the source data as is and defining DWH entries as entire or partial master records. A Hybrid approach (data-less access or zero-latency data warehouses) incorporates components of both schemes. In all approaches, data quality and metadata management are often done parenthetically, and should be executed by a high performance, extremely flexible and easy-to-use data quality system.

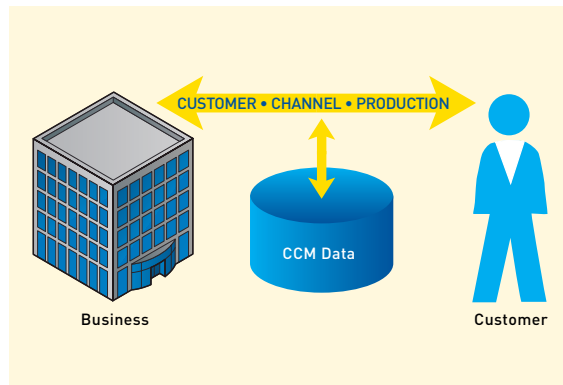
### Data Quality Functions in MDM

A salient feature of a customer-based MDM system is a unique customer identifier that maps all relevant structured customer data. This customer data may include personal information (name, address, phone, email, etc.), business and consumer demographics, communication preferences, order history (product, location, reclamations, satisfaction indicators, etc.), relevant financial support data for credit/underwriting, and contact management history.

Fully implemented MDM is an important feature of Customer Communication Management (CCM). CCM data is broadly categorized into three major groups: Customer Intelligence, Channel Intelligence, and Production Intelligence. Customer Intelligence manages customer data more effectively. Channel Intelligence breaks down delivery channel barriers for more consistent communication. Production Intelligence transforms data for more effective message delivery. Customer Communications Management MDM describes the integration of structured customer data, unstructured content, customer emails, documents, or messages, and structured production control data.

MDM supports closed-loop CCM by relating a customer response (e.g. form, call, e-mail, letter, or purchase transaction) to the initially transmitted communication, thereby enabling intelligent business process management (BPM) processes to partially automate communication transactions.

## MASTER DATA MANAGEMENT PROVIDES THE MEANS TO SUCCESSFULLY AND CONSISTENTLY STORE AND SHARE CORRECT CUSTOMER INFORMATION ACROSS DIVERSE SYSTEMS



*CCM involves recording and utilizing data and metadata about customers, communication channels, and message production.*

Data Quality processes ensure the accuracy, completeness, and consistency of data in not only CCM data structures, but also across entire enterprises and multiple databases, systems, channels, production methods, and customers. This is an increase in scope and role from earlier expectations of data quality management.

### Creating Unique Customer Identities

An important data management task is creating and maintaining a unique, universal customer identity. There are several ways of doing this, most of which involve some sort of hashing method of either indexed or comprised data. For example, a union of partial sets of data, such as last name, address1 field, and zip. This key generation method is arbitrarily expandable to additional fields and has been successfully used to create unique keys for data sets in excess of 100 million unique row entries.

For example, associating geolocation information with addresses provides a method to achieve six sigma level confidence in unique key assignments. Encoding latitude and longitude, zip code centroids, or street centroids are options to extend this scheme.

### Maintaining Customer Data Quality

Once data architects have determined how to create unique customer master keys, they must decide how and where to store master records. Outlined below are a few considerations involved in these decisions.

In a federated system various components of customer information could be designated as master fields and stored in the source systems. A customer master record need not be stored contiguously in a single database. A master record can consist of various fields of reference in various source databases.

An alternative approach is to create a data warehouse that contains copies of customer data. Entries can be designated as master records. The data warehouse must be updated with relevant frequency. Often they are created and maintained by specialized data integration systems that utilize Extract-Transform-Load (ETL) capabilities to move data between sources and targets.

Many Data Integration Systems lack the depth and sophistication in normalizing, consolidating, and matching what a specialized Data Quality Platform offers. In such contexts, the Data Integration System relies on a Data Quality Platform to provide salient information services for data quality and metadata support. A MDM Hub will ensure master customer identities by either storing the data in a repository (data warehouse) or retaining indexes to data in the business operating systems (federated configuration).

In a similar context, a Data Quality Platform can take on the responsibility of providing “trusted” data to the CDI System. The CDI System can use a Data Quality Platform to normalize and correlate data entering the enterprise from outside (or inside, as the case may be) sources. A MDM Hub relies on the Data Quality Platform to ensure correct associations between data and its appropriate

## Data Quality Considerations in Master Data Management Structures

6

master identities. The Data Quality Platform becomes a tool that data management staff can use to continually and effectively ensure operational data quality.

### Customer Data Quality Platform

The Pitney Bowes Business Insight (PBBI) Customer Data Quality Platform (CDQP) is the principal enterprise data quality solution. It is the culmination of a three year research and development effort to build the best-in-class data quality engine. As a Service Oriented Architecture (SOA), the CDQP is a server-based framework that accommodates a wide variety of individual data quality information services. The CDQP can innately be configured as a fault-tolerant, clustered hardware configuration. The administrative console offers complete visibility into transaction types, counts, IP access, and user security. It is implemented in Java and is operable under Windows, Solaris, HP-UX, AIX, Suse and Linux.

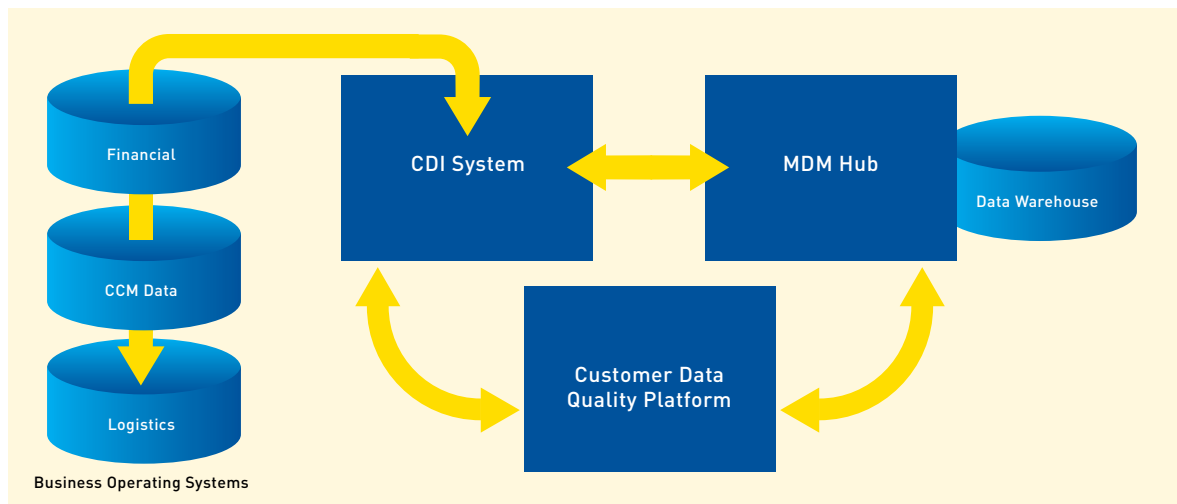
The CDQP is equipped with a wide variety of built-in data quality services including, world-wide postal address standardization, sophisticated geospatial processing, generalized lexical parsing in a variety of languages, and

a sophisticated best-in-class matching service. The CDQP offers built-in Unicode processing and allows the processing of names, addresses, e-mail addresses, telephone numbers, and account numbers – basically any customer data you might encounter.

Variants of the CDQP are implemented in a variety of specialized data quality functions for environments such as SAP, Oracle, and Siebel. The CDQP can be accessed from applications either via an API (.NET, Java, .COM, or C++), JDBC, or WSDL-defined web services.

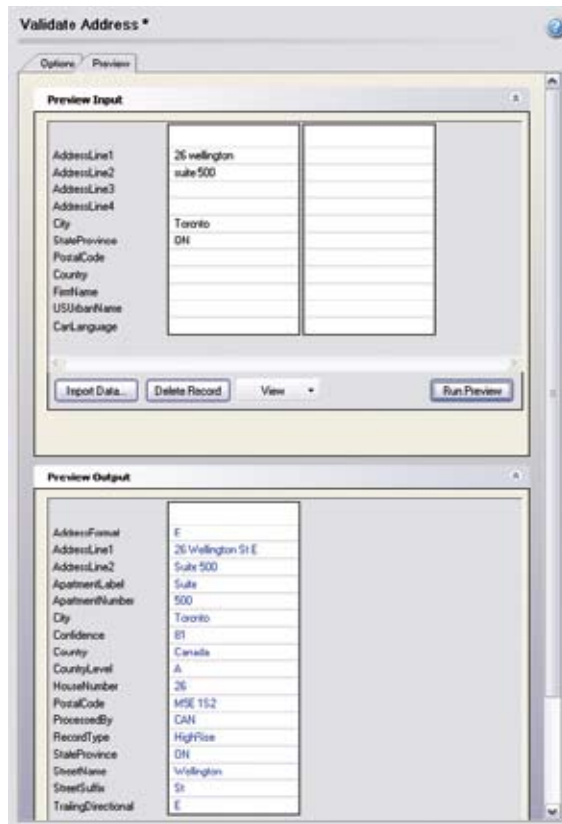
### Master Data Quality

Perhaps the greatest feature of the CDQP lies in its utility to develop and maintain data quality management rules. The CDQP features a graphical rules editor known as Enterprise Designer. Data or system designers can construct data quality processing rules using a rich syntax of parametric data stream operators including, a variety of routers, combiners (merge), sync, and sort. Input and output are implemented either via simple file structures or JDBC drivers. Some Type 4 JDBC drivers are provided, and the CDQP can connect to any available JDBC driver.



A multi-modal platform to ensure customer data quality in a variety of architectural contexts.

## PITNEY BOWES BUSINESS INSIGHT'S CUSTOMER DATA QUALITY PLATFORM (CDQP) IS THE PRINCIPAL ENTERPRISE DATA QUALITY SOLUTION



Address parsing within the CDQP

A built-in set of components is available to perform requisite character string processing, from basic functions (pad, trim, mask, truncate) to advanced transformation rules. These operations can be augmented by applying regular expression and domain dictionaries to extract and/or normalize unstructured data into structured, context relevant values. A resulting dataflow becomes a composite service that is added to a library of published PBBI supplied and user-defined services. Hence, services can be created hierarchically. A user-defined service can be nested within another, facilitating the systematic design of atomic and composite data quality services.

The Enterprise Designer features one-click service publishing. A service can be made available as an enterprise dataflow service in one step. This exposes the service and creates a suitable WSDL entry.

# Data Quality Considerations in Master Data Management Structures

8

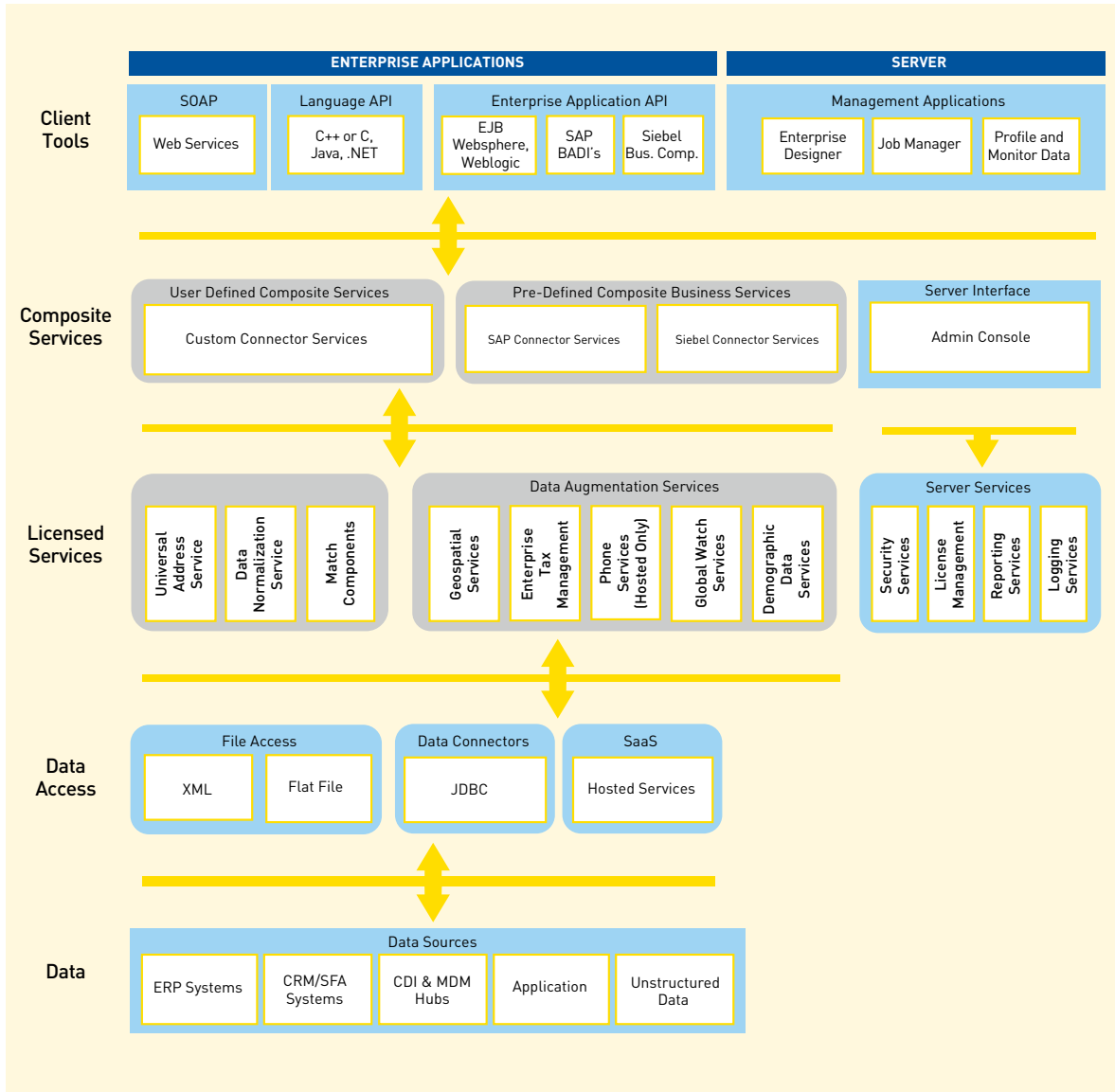
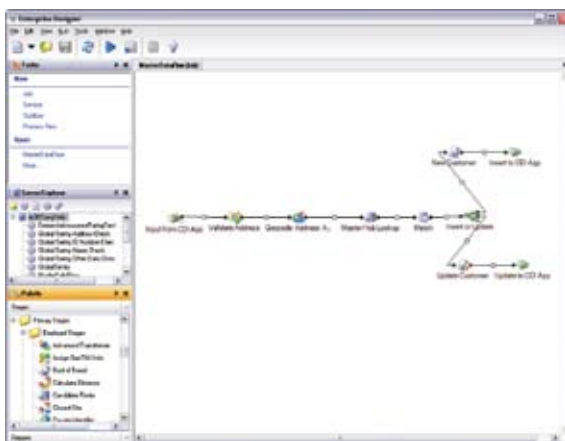


Illustration of the CDQP Architecture

# THE PITNEY BOWES BUSINESS INSIGHT'S CDQP REPRESENTS THE BEST-IN-CLASS DATA QUALITY ENGINE

## Master Data Quality Transaction

The following diagrams illustrate CDQP transactions within the master data quality context.



Customer master data record update transaction example

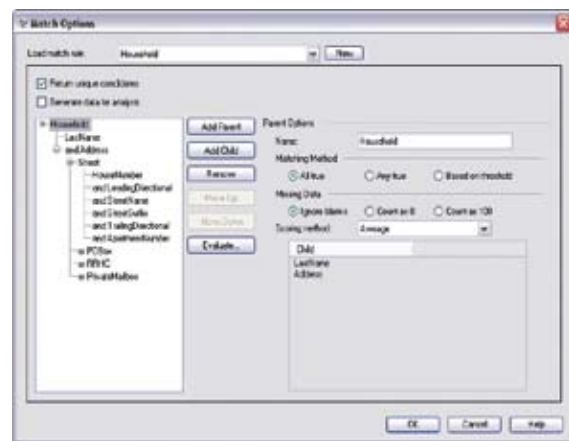
Each node in this flow is parameterized for its appropriate context – settings are stored and retrieved from an internal database. This is just a simple example; real-world flows incorporate built-in exception processing features, multiple synchronous (or asynchronous) flows, and context-appropriate naming.

The flow transactionally receives a message via a web service from an unspecified CDI Application. Subsequent steps validate the address, and add a location geocode. The subsequent two stages interact with the MDM Hub system. The first component looks up the address in the existing population of records. The second (Matcher) then attempts to match against the retrieved set based on data agnostic, probability based match rules, and algorithms. The parametric output then either creates a new customer or updates an existing customer record within the MDM Hub. The master record is also provided to the CDI Hub for optional insertion back into the source system.

As stated previously, Enterprise Designer flows can be run

in either batch or transactional mode. In reality, particular algorithmic sequences are more appropriate for large file processing, but the CDQP itself makes no differentiation.

9



The CDQP is supported by a sequenced, powerful matching engine with sophisticated built-in and extensible matching features.

The Matcher provides built-in algorithms such as keyboard operations, metaphone, soundex, edit distance, and country specific phonetic algorithms. The Matcher offers a three-stage, hierarchical rule specification capability that includes weight scores, cross-field matching, and conditional matching across any number of fields and any number of algorithms per field. The example above shows the built-in house-holding match rule that supports the detection of individuals as part of a household.

The CDQP offers a variety of built-in name parsing and name variation knowledge bases with cultural context to account for variations created through the process of transliteration from original language script to Romanized form. For example, an Arabic script for Mohammed when transliterated into its Romanized form can generate over fifty variations depending on the region and particular cultural nuances. Parsers can also be extended to parse any field for purposes of normalizing encoded text and subsequent match input.

## Data Quality Considerations in Master Data Management Structures

10

### Summary

Correct customer information is the life blood of every company. While gathering and storing vast amounts of customer data is a worthy task, it is more important to properly combine, manage, and share this information within the company and across the enterprise. Master Data Management provides the means to successfully and consistently store and share this information across diverse systems, thereby ensuring a properly managed and centralized corporate data structure.

Pitney Bowes Business Insight has over 20 years of commitment to improving communication and customer data management efficiency. The CDQP development team has years of aggregate experience dealing with real-world data quality issues. This experience, combined with emergent web service standards, service oriented architectures, and commonplace distributed customer data, has resulted in strategies inherent to contemporary Data Quality.

We invite system and data architects to consider the utility of common customer data quality systems based on the CDQP when preparing to implement a Master Data Management solution.

# PITNEY BOWES BUSINESS INSIGHT: YOUR SOURCE FOR ANSWERS AND SOLUTIONS

## About Pitney Bowes Business Insight

11

Operating as one division, Pitney Bowes Group 1 Software and Pitney Bowes MapInfo are now called Pitney Bowes Business Insight. Pitney Bowes Business Insight offers a unique combination of location and communication intelligence software, data and services that can be used throughout an organization.

We combine our deep industry knowledge with our strategic analysis offerings and apply our expertise to help you take action that leads to better, more insightful decisions. You will get a more accurate view of your customers, and integrate that intelligence into your daily business operations to increase revenue, improve profitability and enhance operational efficiency.

For more information about the Customer Data Quality Platform and/or Pitney Bowes Business Insight call 800-327-8627 or visit [www.pbinsight.com](http://www.pbinsight.com).



#### UNITED STATES

One Global View  
Troy, NY 12180-8399  
main: 518.285.6000  
1.800.327.8627  
fax: 518.285.6070  
[www.pbinsight.com](http://www.pbinsight.com)

#### CANADA

26 Wellington Street East  
Suite 500  
Toronto, Ontario  
M5E 1S2  
main: 416.594.5200  
fax: 416.594.5201  
[www.pbinsight.com](http://www.pbinsight.com)