Name and Address Data Quality

by Sophie Allen

MasterSoft Research Pty Ltd, Sydney Australia

1. Turning Customer Data into Information

In today's competitive business environment corporations in the private sector are being driven to focus on their customers in order to maintain and expand their market share. In the public sector the de-regulation and privatization descending upon government departments is also pushing these traditional monopolies into competition. Suddenly, telephone and power utilities are scrambling to discover their customers. This shift is resulting in customer data and information about customers being viewed as a corporate asset. As with any asset, it has certain attributes and value: Who are our customers? What is the profile of our customers? What are our customers individual needs? What is the profile of our most profitable customers? To be able to accurately answer these questions, an organization must be equipped to analyze their customer data, and use the data to derive information.

Historically, customer data has been gathered by separate business units within an organization to fulfill government regulations or as a means to send a piece of mail. This data is generally not of sufficient quality to derive answers to the types of questions organizations want to answer. Information technology (IT) departments are therefore being asked to delve into this new data management arena. They are setting out to improve customer data quality and effectively manage customer data. The challenges faced to achieve these goals are the ability to:

- attain consensus among all lines of business to participate in data quality projects where ownership of customers is still along lines-of-business;
- re-align the organizations structure to be customer focused;
- cost-justify projects that are corporate wide and whose benefits are difficult to quantify; and
- acquire skills in data management to minimize risk and not re-invent the wheel.

2. Dealing with Name & Address Data Quality

Customer data consists of many components with Name & Address being the key. A person or companies Name & Address is THE identifier of a customer. Organizations may use: policy, bank account, or even telephone numbers but these really only apply to a particular line of the business. In telecommunications, for example, a customer may have a series of numbers: mobile, residential, business, calling card etc.

Name & Address as a data type is very difficult to deal with and manage. This data is volatile... customers come and go, addresses change, names change. This data is often cluttered when entered. Name & Address fields on data entry screens are usually free format and ripe for use to enter comments without any edits. Name & Address is subjective... it can be written in a number of different ways and still be the same. Finally, there is no application independent standard for Name & Address that can be used as a measure of quality.

Name & Address data quality needs to be measured at two levels: the quality of *individual* Name & Address components, and the integrity of the customer base as a *whole.* At the *individual* level metrics include the degree of: structure, content, completeness, consistency and accuracy. While the integrity of the *whole* customer base can be measured by examining the accuracy and completeness of relationships identified: between individual customers and between customers and the organization.

3. Name and Address Data Integrity Software

The Name and Address Data Integrity Software, (NADIS) is a suite of software products developed by MasterSoft Research Pty Ltd in Sydney, Australia. It was specifically developed to assist organizations in the management of Name and Address. Two expert systems which parse and match data are at the core of the software. These rule-based systems or engines are implemented with transition table technology. The engines operate on a variety of platforms and are packaged into three products: ScrubMaster, SearchMaster and Onlooker. ScrubMaster is used parse files of Names and Addresses. SearchMaster is used to identify relationships within and across files. Onlooker incorporates both engines for the on-going protection and management of clean databases.

The parsing engine converts Names & Addresses to a common, structured and standardized format. The application independent common format developed by MasterSoft is called Universal Name and Address (UNA). The parsing engine also reports on and corrects data inconsistencies. Among the data inconsistencies identified are missing or extraneous data, invalid characters or illogical information.

The matching engine identifies relationships between parsed Names and Addresses. It emulates the same reasoning we use when comparing the differences and similarities between Names and Addresses. The matching engine accounts for the common typographical, transliteration, and elision errors made when keying data.

1.1 Pressure to focus on the Customer

Corporations are becoming customer focused because of competition. In the private sector, competitive pressure is forcing corporations to know and get closer to their customers. They are targeting channel marketing efforts to the individual rather than rely on mass marketing mechanisms. This same shift is emerging in the public sector where the de-regulation and privatization of traditional government monopolies is pushing these organizations to discover their customers.

At a strategic level corporations are seeking to develop the corporate - customer relationship to increase revenue while minimizing the cost of sales, and to maintain or increase market share. At an operational level, these strategies are being implemented with the following objectives:

- Improve customer service;
- Provide better information for decision support;
- Improve customer retention;
- Improve risk assessment and fraud detection; and
- Increase sales by cross-marketing products and services to the customer base.

Take a telecommunication company's initiative to improve customer service by setting up a one-stop service center. A customer will be able to ring one main number and the customer service agent is equipped to handle ALL the customer's queries. The customer service agent must have a picture of all the customer's dealings with the corporation. The ability to provide this level of service hinges on having customer focused data.

1.2 What does this mean in relation to customer data ?

Traditionally, organizations have used internal numbers to relate with, and identify, customers. From an IT point of view numbers were great unique identifiers of items such as insurance policies, bank accounts, passports, trust accounts, and transactions. Product systems and lines-of-business were built around these numbers. Of course, all interaction between the customer and the company revolves around the number. The problem is that the same customer may have a series of identifiers. In most systems, a user could not search the database to find a customer by their name.

Data about the customer was only obtained if the organization needed to send mail to the customer or for reporting requirements to government departments. In a bank a mailing address for a credit card is necessary to send a statement where as a passbook account may not need an address. This type of data was not the focus of the system, but merely collected and maintained.

MasterSoft has assisted many organizations in making the shift from having customer data to having customer information. These customer data quality projects typically start with the following questions:

- How many customers does the company have?

Typically, a line of business can enumerate the number of policies, accounts, agents but it can only *guess* how many customers that translates to. Expanding the scope across the corporation usually yields a shrug of the shoulders.

- What attributes are used to identify a customer? or simply Who is a customer?

Agreement on who is a customer is another difficult proposition. If a customer last bought a product or service 2 years ago, are they still a customer? Is the name in a care of address a customer? Are agents who act on behalf of others customers? Is the beneficiary of a policy a customer?

- What is the profile of the company's most profitable customers?

Which of all the customers are the most profitable? What areas do they live or work in? How many households does this comprise? How many of these customers live in the same household?

Before this information can be extracted from the customer base, the organization must measure and improve the quality of its customer data.

1.3 Challenges in turning customer data to information

Turning the customer data in these legacy systems into customer information as the focal point of systems poses significant challenges to the enterprise:

Agreement from Lines of Business - To fully exploit customer as a corporate asset, all lines-of-business need to participate and contribute to a strategy. This commitment is frequently difficult to obtain because each line of business has its own goals and accountabilities. For customer focus to be profitable it must be embraced by all levels of the organization and initiated by the very top.

Organization Structure - Few organizations, have a structure chart with the title *Client Data Manager, Client Data Owner or Data Custodian*. The enterprise must recognize and allocate these distinct responsibilities. With systems built along product lines or lines-of-business, these three titles (if they existed) were within that area. If client data is a corporate asset, resolving client data ownership and management is not an insignificant task. A recent survey conducted by our North American distributor, Group 1 Software, reinforced this point. Of the 100 responses from senior IT executives, there where 50 different answers to the question who owns the data ?

Cost-Justification - Which departmental budget absorbs the costs to measure, improve and maintain customer data quality to meet the operational and strategic objectives? It is difficult to tangibly cost-justify the direct value of this corporate asset, and data quality projects need resources from across the company.

1.4 Challenges faced by Information Technology Departments

One common element is that IT departments are charged with the task of physically providing information, data quality is a new arena. Over the years, IT departments concentrated their efforts in the 70ís on hardware, in the 80ís on software and in the 90ís are moving towards data.
Even today, agendas of many Data Warehouse conferences do not include seminars on Data Quality. Topics include meta data, databases, data warehouse architecture and technology, decision support systems, principles. The EXTRACT, TRANSFORM and LOAD phase of the process seems to be always glossed over as if it just happens. This phase consumes a lot of effort and is not a sexy proposition. However, failure to appropriately address data quality issues will certainly guarantee the failure of Data Warehouse projects.
The skills required to effectively manage data quality are typically not available in IT departments. Most application development project teams include project managers, analysts, programmers, and business analysts. It is rare that skills in managing data for the corporation are included in the project team. As a software development house, MasterSoft finds it difficult to locate staff with experience in data quality. It is difficult to hire people who know answers to questions such as: What are the issues involved? What are the pitfalls? Where do we start? How and what do we measure? How do we go about all of this on a scale of millions of records?

2.1 Why is Name & Address difficult?

Name & Address is the data which is the only common identifier of a customer. Even though it is key to identifying every customer, it is subjective and volatile and therefore difficult to manage. This view was supported in the same survey conducted by Group 1 Software in which Name & Address had the highest response (31%) to the following question, Of all the legacy customer data you've collected, what type of data presents the most problems when it comes to ensuring its integrity?
The properties which make Name & Address difficult include:
Subjective - People can write the same Name & Address data in many different ways. When we interpret data, we account for differences by applying knowledge about known cultural norms in the way Name & Address data is written. This is easy when a person physically reviews and

compares two names and addresses but it becomes an onerous task when having to automatically apply these norms to millions of Names & Addresses. The problem is further compounded by the different ethnic backgrounds of data in a global market.

Free Form- Name & Address fields in data entry screens are typically not structured. Being perhaps the only free form fields available, they are ripe for entry of comments, account numbers, statuses etc. Usually, anything that doesn't fit into another field is jammed into Name & Address fields. Also, Name & Address is very difficult to edit for completeness, consistency or accuracy. So, Name and Address is just added or updated regardless of quality.

Volatile - Names & Addresses change all the time; they are not static. This volatility is caused by corporate mergers, marriage name changes, movement of residential addresses, individual customer preferences - in communication, privacy of activity, changing residential address, changing postal addresses etc. This volatility needs to be managed.

Standards - Using conformance to a data standard as a measure of quality is difficult because there is no independent data standard for Name & Address either internal or external. Historical use of Name & Address has been for particular purposes - mailing or government requirements. A company conforms to these standards because of enforcement by the receiving agency. For example, when a bill mailing run is performed, data is extracted from the database and massaged (if possible) to meet the local mailing standards before addresses are affixed to the envelopes. The massaged customer data held by the organization is typically not updated. The same data may be extracted and massaged differently to meet another standard. The quality of the data held by the organization is not improved.

2.2 Definition of Customer Data Quality

A definition of customer data quality that MasterSoft has developed based on experience with our customer base is*: The extent to which customer information remains reliable and consistent across the corporation .* Unfortunately, most corporations seem to view data redundancy or duplicate records as a measure of Name & Address quality. The reality is that this is just one facet of customer information quality. Name & Address Data quality needs to be measured at two levels:

- the quality of **individual** Names & Addresses
- the integrity of Names & Addresses as a **whole**

Possessing high quality individual Name & Address data is an essential foundation to deriving reliable information about customers.

2.3 Quality of Individual Names & Addresses

Before an organization can extract reliable information from Name & Address data as a whole, it needs to identify, validate and measure the quality of the individual Names & Addresses. This is the necessary first step. The following facets of Name & Address can be used to measure the quality of *individual* Names & Addresses across the entire customer base.

Structure - Does the Name & Address have the appropriate structure necessary to meet all the organization's needs? Is the data completely structured? Are the key elements structured? Free format data does not allow the organization to effectively communicate with the customer in a wide variety of ways or perform market segmentation.

Completeness - What are the key elements for minimum name data? Is MR JONES acceptable? Will it meet the organization's needs? What are the key elements for minimum address data required? Is an address NEW YORK, NY acceptable? Will it meet the organization's needs? What other elements are missing? Determining the missing elements, can only be undertaken if the data is structured. If the data is structured and the elements are missing, are they required?

Consistency - Has the same spelling been used to reflect the same meaning? Does the Name & Address conform to determined standard spellings for commonly used terms? Do a customer's initials appear on one system but a full name on another?

Accuracy - Is the data accurate? Is the name spelt correctly? Is the address current, or has the person moved? Is the Name & Address reliable?

2.4 Integrity of Name & Address as a whole

Once the quality of the individual Names & Addresses is measured and improved to the level required, the Names & Addresses can be used to **reliably** :

- Remove duplicates: Identify how many customers and who they are without clutter.
- Derive other data by verifying or allocating: telephone numbers, geocodes, company codes etc using external sources.
- Provide information: Which are our most profitable customers?

MasterSoft's experience with these three objectives is to attack them in the order listed. Duplicates need to be identified or removed, otherwise effort is wasted on deriving other data for Names & Addresses which may be eliminated. Once these steps are completed, the data can be used to extract reliable information.

The information required of the customer data typically falls into two categories of relationships or links:

Between a customer and the organization- For a particular customer or group of customers it is important to have a view of all dealings the customer has with the organization as a whole. A bank, for example, may want to determine all banking products a customer has (either solely or jointly).

Between a customer and other customers- For a particular customer or group of customers it is important to have a view of all relationships the customer has with other customers. Examples in banking include:

1. In order to measure exposure risk to a particular corporate client, the bank must be able to identify all the divisional, departmental, and different offices of same company.
2. Knowing the make-up of a household allows the bank to cross-market products effectively.

The quality of this derived information can be assessed based on:

- Completeness - Are there any *real* relationships that exist in the customer base that were not identified?
- Accuracy- Are there any identified that really don't belong or are incorrect?

2.4 On-going Client Data Management

Customer data quality projects do not end once the quality is assessed and improved. They also do not end when information is derived. To maintain the quality and the information, the organization must implement an on-going strategy for customer data management.

3.1 About MasterSoft

MasterSoft International was founded in Sydney in 1989 to develop knowledge-based Name & Address matching software for the Australian market. The company has expanded operations into New Zealand, and appointed distributors in the USA, and Canada. To further increase its level of customer service MasterSoft International divided its operations in 1995 to form MasterSoft Research Pty Ltd, also headquartered in Sydney. MasterSoft Research is solely responsible for developing NADIS (Name and Address Data Integrity Software) products for vertical and global markets and to undertake joint research projects. Over the last six years MasterSoft has grown dramatically and is now generating revenue of over AUD $6 million with 22 employees. NADIS is used world wide by over 70 customers ranging from government departments to large insurance and banking corporations. Two case studies of NADIS customers are included in the appendix which describe how NADIS has been used to manage Name and Address Data Quality.

3.2 NADIS Products

NADIS incorporates MasterSoft's Name & Address processing products, ScrubMaster, SearchMaster and Onlooker. These products use sophisticated inference engines, which emulate the processes used by people to assess names and addresses, locate relationships between client records despite keying errors, spelling discrepancies and differences in data format.

ScrubMaster: structures and standardizes a company's files of Name & Address data. The software identifies every element in a Name & Address record, and then converts these elements to the Universal Name and Address (UNA) data standard. UNA is NADIS' standard format for describing Name & Address data. The UNA is critical to customer information system (CIS) applications for two reasons. First, it provides a uniform standard so users can effectively manage and model their customer Name & Address data. Second, it enables users to easily move "cleaned" data between different computer systems and applications.

SearchMaster: matches files of Name & Address data. The software processes the output of ScrubMaster to analyze all relationships between Name & Address records. It emulates the inferences used by people, accounting for such variables as spelling and phonetic differences to provide reliable and accurate matches on name or address, or Name & Address combined.

The result is much cleaner names and addresses because the information has undergone much tougher scrutiny than less sophisticated software can provide.

Onlooker: is an on-line package combining Name & Address scrubbing, matching and search key generation. It can be installed as a batch-driven system or as a true interactive system. Onlooker uses the same scrub and match engines used in other MasterSoft products.

Competition is forcing corporations to get to know their customers. This poses a number of challenges for corporations and their IT departments to switch from customer data to customer information. IT departments are now focusing on data quality which is a new arena,

Name & Address is the key data component of customer data yet because of its nature it is difficult to manage. Frequently, data quality of Name & Address is mistakenly equated to duplication. To properly assess Name & Address Data Quality, a corporation needs to identify and measure the quality of individual Names & Addresses before the integrity of the Name & Address data as a whole. This approach will provide the corporation with the reliable information it requires about its customers to achieve related business objectives.

NADIS, a suite of Name & Address Data Integrity products developed by MasterSoft Research, is used by organizations world wide to measure and improve the quality of names and addresses of customers.

This case study is taken from a DEETís (Australian Federal Department of Education, Employment and Training) presentation at the MasterSoft User Group in 1995. The presentation was prepared by:

- *Carole Hanstede and Simon Fisher (User Support and Administration)*
- *John Blazko and Robert Heggie (Technical Development and Support)*

Background

Australia-wide the Department of Education, Employment and Training has over 16,000 staff and 676 offices administering a wide range of funding programs that are supported mainly by computerized application systems and occasionally by clerical support. The Department's six million plus clients include jobseekers, labour market program participants, Australian and over seas students, educational institutions, employers and other commonwealth and state government departments. There are 10 major IT systems and a further 70 smaller systems at DEET running on a diverse range of hardware and software platforms. In the past few years, interfaces have been created between all these systems.

Why Quality of Customer Information is important to DEET?

An ongoing need to support new government initiatives means DEETís applications are constantly being upgraded. Because of the complexity of the hardware and inflexibility of many of the applications, heavy demands are placed on system maintenance and development. This has created a number of problems including duplicated client data. The large number of duplicate clients on DEETís systems has resulted in:

- Lack of data integrity
- Inadequate data matching
- Risk of fraud
- Potential duplication of payments
- Decreased customer service

DEETís battle with duplicate data

In mid 1993, a massive clean up exercise was carried out to remove duplicates. This succeeded in reducing the number of duplicates by over 60 percent. But within 6 months the amount of duplicates rose again to the original level. The department looked for a more constructive and permanent solution, such as putting in place preventative measures at the points of data entry. It was not possible to plug all data entry points simultaneously in 80 plus systems. Since reducing the risk of fraud and duplicate payments is a high priority within the department, they first focused on its financial system.

The seven systems used to make DEET payments hold over 900,000 client records. Prior to introducing preventative measures across these seven systems, DEET developed a common standard for capturing and storing the data. Before implementing preventative action, DEET created a weekly average of 560 duplicates. This compares to 75 duplicates per week which can be easily detected an removed after the on-line preventative measure was implemented.

Future strategies for Client Data

Among DEETís future plans:

- Improve and maintain data integrity in the Student Assistance system and Employment Schemes
- Create a central repository for all DEET client records to be used by all applications
- Investigate the use of laptops to enter client details in remove areas
- Introduce client tracking as most clients register more than once

This case study was provided by MasterSoft for incorporation into a Data Quality Report co-authored by the Cambridge Research Group and Cambridge Market Intelligence. It covers a relatively unique application of Data Quality to help solve common global law enforcement problems.

Background

Formed by the Australian Federal Government in 1989, the Australian Transaction Reports and Analysis Center (AUSTRAC) assists in the deterrence and detection of tax evasion and criminal activity, money laundering from organized and corporate crime, and drug trafficking.

AUSTRAC collects, analyzes, monitors and disseminates transaction report information from domestic and international financial institutions and other cash dealers, each of which is required to report significant cash transactions. The types of organizations range for small-time bookmakers to the 5,500 branches of the four major Australian banks. The latter contribute four fifths of the reporting volumes. Reports produced by AUSTRAC are made available to the Australian Taxation Office (internal revenue) and law enforcement agencies. The Center's annual budget is around US $7.5 million, of which 60% goes to IT.

Objectives / Challenges

Over 7,000 organizations report electronically to AUSTRAC. Each of these has its own format for Name & Address data, so a key issue for AUSTRAC was how to cost-effectively and rapidly standardize and monitor this incoming data. Because of the sheer variety of formats the challenge was to find an automated system solution which could consolidate and de-duplicated this demographic data.

Solution / System Description

AUSTRAC requirements for Data Quality management tools:

- accept Name & Address data in a wide variety of formats
- identify multiple entities in each transaction (for example, sending and receiving parties, joint accounts)
- identify existing details of the parties on the database and link financial transactions to provide relationship information for future investigation
- perform standardization of data into a common format for *further* processing
- ability to integrate into on-line systems
- support flexible enquires necessary for their investigations
- complement existing search key mechanisms with matching
- suitable for their current technical environment, and architected to allow portability to future environments

To use their own words, a solution that offered AUSTRAC a balance of completeness, accuracy, adequacy and timeliness .
The system was developed for a Tandem Cyclone, running non-Stop SQL under Guardian. The AUSTRAC application itself was written in COBOL , and the overall database contains more than 5 million reports, with annual growth expected at 3 million reports. Including the AUSTRAC in-house user base about 1500 tax and law enforcement officers have on-line access to AUSTRAC data.

Benefits / Productivity Gains / Savings

The reports provided to AUSTRAC have yielded significant results in the short time that the agency has been operational. The use of a specialized Data Quality tool has improved the solution in three main ways:

1. Fast development time compared with full in-house coding (the system went live in very short development and integration cycles)

2. Specialized and robust data quality functionality and capabilities are built-in
3. Bottom line results: many suspect transactions have been identified such as:
    - Suspicious transactions were reported by a bank concerning unusual use of negotiable instruments. This information matched with information already held by a law enforcement agency and has led to a major money laundering inquiry. Company directors have now been charged with offences, and substantial revenue is expected to be raised as a result of tax investigations.
    - A pattern of money flows over a period of time pointed to links among suspected organized crime and international drug trafficking activities. This linking of money flows, groups and individuals has generated a new understanding of the way in which organized crime operated in Australia. It has also produced a number of new leads which are contributing to the completion of prosecution cases. Most importantly, this information has given law enforcers a powerful resort to establish co-operation with relevant over seas law enforcement agencies.

Conclusions

The data processing requirements of AUSTRAC reflect those of many commercial systems, but far exceed them in complexity and variety. The ability of the NADIS product to satisfy the AUSTRAC requirements was an example of wise management decision to purchase a product that embeds a high degree of specialization and expertise. Although AUSTRAC probably did not think in terms of building a data quality system per se, without a data quality tool their requirements would have been very difficult to meet so quickly and the project itself could easily have become unmanageable.