APPLYING NAME KNOWLEDGE TO NAME QUALITY ASSESSMENT

(Practice-Oriented)

Kimberly Hess
Acxiom Corporation, Little Rock, Arkansas
Kimberly.Hess@acxiom.com

John R. Talburt
Acxiom Corporation, Little Rock, Arkansas
John.Talburt@acxiom.com

Abstract: Data quality is an essential component for successful Customer Relationship Management (CRM). This paper discusses some of the considerations in the approach, design, and administration of tools and products intended to assess the quality of consumer names.

Keywords: data quality, TDQM, data quality tool, name quality assessment, name knowledge

INTRODUCTION

Corporations worldwide are noticing the business advantages that emerge when they integrate a strategic data quality initiative into their overall mission such as a Total Data Quality Management (TDQM) system [2]. Innovative technologies that use multiple stores of data across various points in a corporation help create a 360° view of customers, and this has led to an unprecedented need for high data quality measures and procedures.

Although corporations maintain many pieces of data about a consumer, the foundational elements of the consumer's address often times are the road blocks to successful matching, and successful matching is needed for cost-effective data integration. Once corporations determine the delivery point information (the address), they may assume the remaining data is the consumer name. The name component(s) of the data, however, should also be analyzed to help facilitate high-quality data integration. Using expert naming tools prior to applying data integration algorithms helps corporations integrate a successful data quality initiative.

BACKGROUND

When the Customer Data Management Industry was primarily focused on Direct Mail Marketing (DMM), postal delivery point validation, cleaning, and correction were elevated to a fine art, but little attention was paid to the name component – a mail piece addressed to "Resident, 123 Oak St." was just as

deliverable as one sent to "John Smith, 123 Oak St. Now that the current emphasis is on Customer Relationship Management (CRM) with its dependence on Customer Data Integration (CDI), the situation is very different. The name component of an address has become just as important as the delivery point component. Unfortunately, the state-of-the-art for name quality assessment and cleaning is not as advanced; far fewer tools and resources exist for the name component than for the delivery point component.

As businesses continue to use multiple channels of customer contact, the need for higher quality CDI solutions has increased. A customer's name and delivery point are the most commonly encountered items for customer identification. The application of name knowledge to name quality assessment is a key component in the critical path for improving the quality of CDI and customer recognition processing.

The universe of postal delivery points is well defined, at least for the U.S. Address hygiene and validation applications incorporate standardized delivery point components defined for addresses with data such as the U.S. Postal Service files. When used in matching, these files can drastically increase the confidence of deliverability. Significant knowledge about the deliverability of a delivery point is leveraged, and the deliverability in DMM applications is essential to minimizing cost. Combining the standardized delivery point components for an address with data validated for these components creates a powerful tool for the DMM Industry.

Unlike delivery points, names for individuals and businesses can be created at will. These names cannot be derived from a comprehensive, finite source. Consequently, the assessment of a name as valid or invalid carries a higher degree of uncertainty. Assessing name quality is a more difficult process that requires collecting and cross-referencing name knowledge from a variety of sources, and properly applying that knowledge in the context of user needs.

Name Knowledge Resources

Mining Internal Best Practices

In a large data management company, many one-off solutions may be scattered throughout different channels to identify a number of similar, if not identical, validation rules for name quality in particular solutions. Consolidating these efforts requires focus and could be a significant undertaking. New solutions development and utilization of consolidated data expertise, however, become a resource to the entire corporation. Furthermore error reduction in name data that is incorporated with the matching mechanism in a CDI solution can actually *decrease* the functionality necessary for matching.

Data Resources

Several methods exist for gathering data and driving name assessments. For example, registered names in databases such as the U.S. Census, Social Security Death Index, Fortune 500 Business List, and USPS standard business abbreviations are all sources often used in gathering name data. In addition, various Internet sources maintain lists for vulgar words, phrases, and suspicious names. Factual and verifiable names are the highest priority, and names provided from corporate-specific solutions, which vary from customer to customer, should be incorporated only after research across multiple verifiable sources. Frequencies associated with names at the first name and last name level, or distinguishing the usage of first and last names separately, can increase the assessment depth of a name tool.

Design Considerations

Performance Efficiency versus Comprehensiveness

The approach for assessing most data problems includes determining how to assess the majority of the

data without elaborate analysis. Narrowing the overall compilation of sources to specific cases that are considered less than common, and then applying a more complex analysis to the data subset proves very beneficial in ensuring that computational time is wisely spent (i.e., less obvious data issues for the highest quality results).

Optimistic versus Pessimistic View

To some extent, the current operator is somewhat "optimistic" because it primarily looks for patterns or conditions that might be a result of accidental data entry or inadvertent process errors (i.e., honest mistakes). However for some files such as self-reported name and address data from the Internet, the errors can be more intentional and harder to detect. Including more pessimistic rules to handle this data can have the undesirable side effect of substantially increasing processing time. Also, in some cases, correct name data may be flagged as an exception (e.g., celebrity names).

General-Purpose versus Customization

A name assessment system should appeal to users across a corporation because it covers a spectrum of needs from the most basic name issues to the more complex. One consideration is to ensure that the results can be easily interpreted for a specific application. Another is to provide record-level details that allow users to apply specific business rules after the analysis. Additionally, users need a clear understanding of basic assumptions and approaches of name assessment for CDI processing. Examples would be whether the system should attempt to classify all input names or assume names are valid unless an exception condition is met. Finally, multiple name acquisition points, such as the Internet versus a call center, offer insight into potential options necessary for users. Diverse data points handling will be discussed at length in a following section focused on optimistic versus pessimistic approaches.

An Example System

Some of the choices and compromises described in this paper are exemplified by a system called "NameCheck". NameCheck is a rule-based expert system that analyzes U.S. consumer names for specific patterns of words, characters, and symbols that may indicate a data quality issue. The following is a discussion of where NameCheck falls within the spectra of the three considerations of the previous section.

The NameCheck design definitely opts for efficiency versus comprehensiveness. Even though NameCheck currently runs in a highly parallel "pipeline" system where each records passes through multiple transformation operators, execution time is still an issue. In a production environment where file size routinely exceed 100 million records, small increases or decreases in efficiency are magnified in terms of overall run time. For this reason, NameCheck uses relatively small reference tables, approximately 40,000 entries, relying primarily on analyzing patterns built from these tables to discover name anomalies. The reference tables comprise name components, such as first name and last name, associated with specific knowledge used in assessing name quality. The graph shown below illustrates the effect on efficiency as the size of reference tables increase.

The graph below illustrates the impact that reference table size has on performance time:

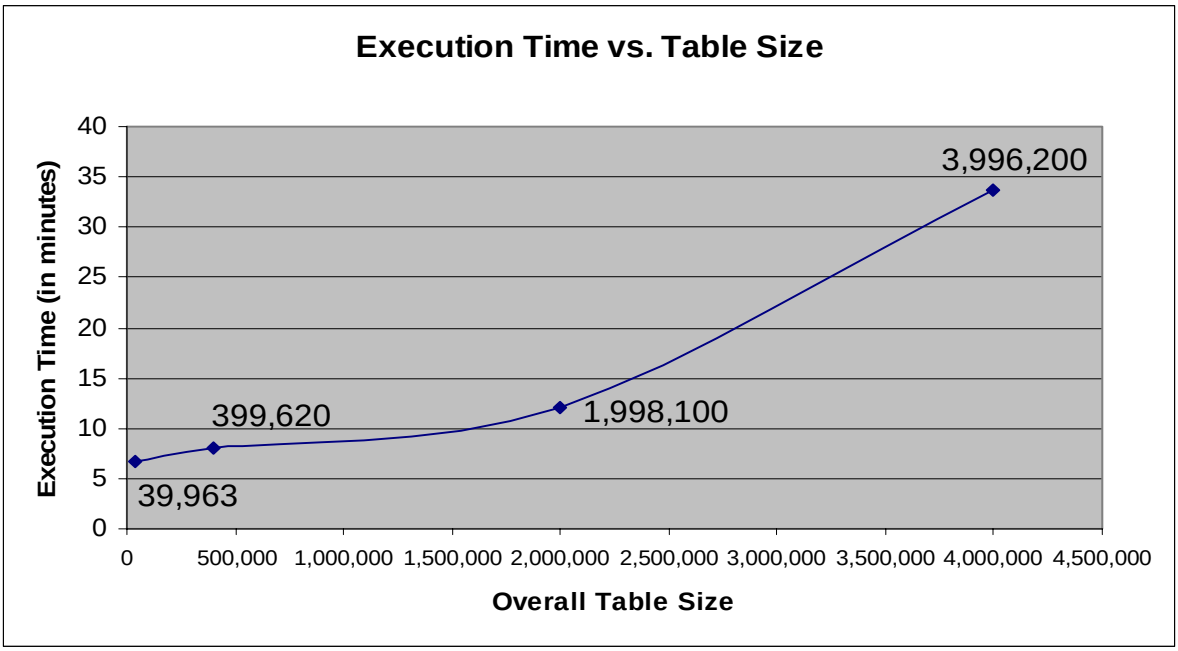

Figure 1: Impact of Reference Table Size on Performance Time

Considering NameCheck's primary analysis focus on patterns, the system is considered optimistic. It assumes the name to be valid unless it detects a specific pattern indicating otherwise. Consequently, NameCheck achieves optimal assessment on files that contain non-intentional transcription and processing errors versus files with large numbers of intentionally deceptive names. This is due in large part to the difficulty of detecting suspicious names, such as "Ben Dover" [3], by pattern determination referencing tables and identifying all attributes of the name rather than depending solely on a table lookup. Identification of multiple names, such as "Mr. and Mrs. John Talburt", and names missing a surname, such as "Diane Kathryn" are examples of patterns where the NameCheck assessment performs with "optimistic" data types.

Finally, NameCheck is general-purpose system. The current implementation does not allow users to supplement the basic rules or tables of the system to fit specific application. However, it is comprehensive enough for corporate-wide use and allows for multiple input options. As some compensation, the system will write specific reason (pattern) codes into the record along with providing a hygiene name, when applicable, to leverage further analysis and decisions in down stream processes. At-a-glance, using the NameCheck report, users can determine the overall quality of their file with summary statistics and examples for all reasons codes. Other information available from reporting includes a breakdown of the names comparable to census statistics.

The figure below illustrates an overall example system design:

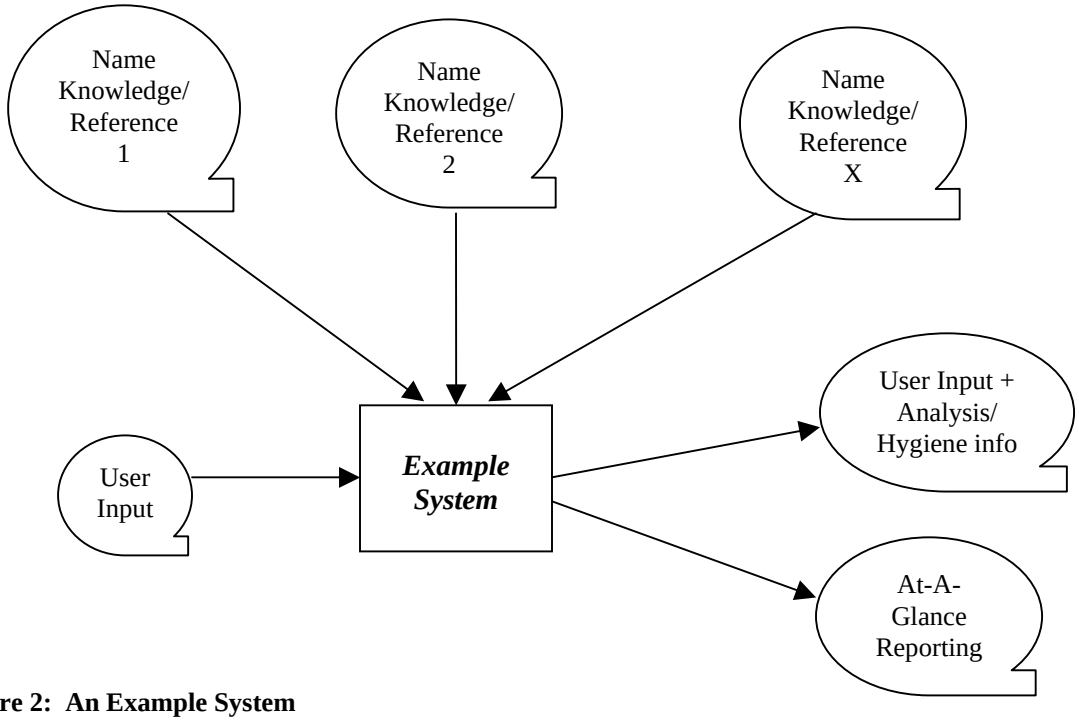

Figure 2: An Example System

The figures below illustrate multiple processing options available to users:

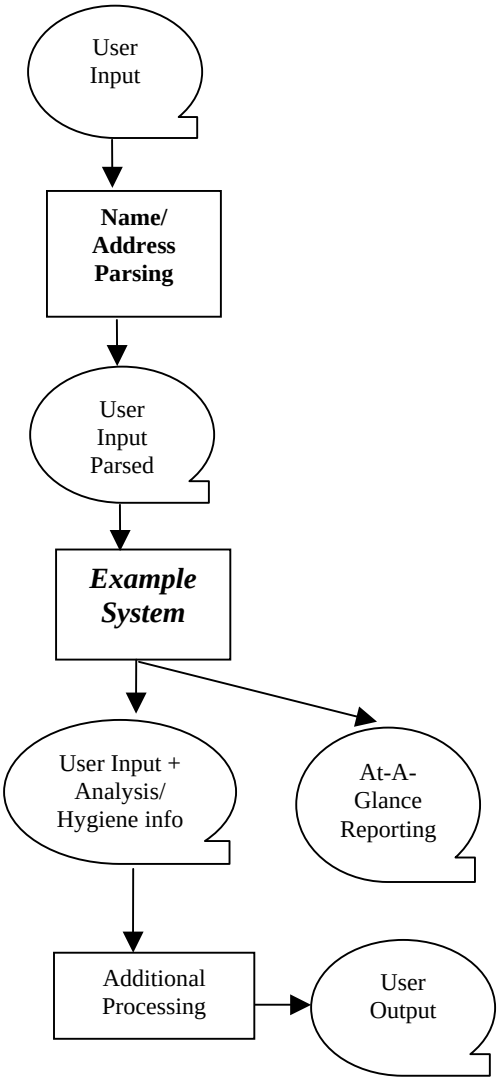

Figure 3: User Input, Parsed

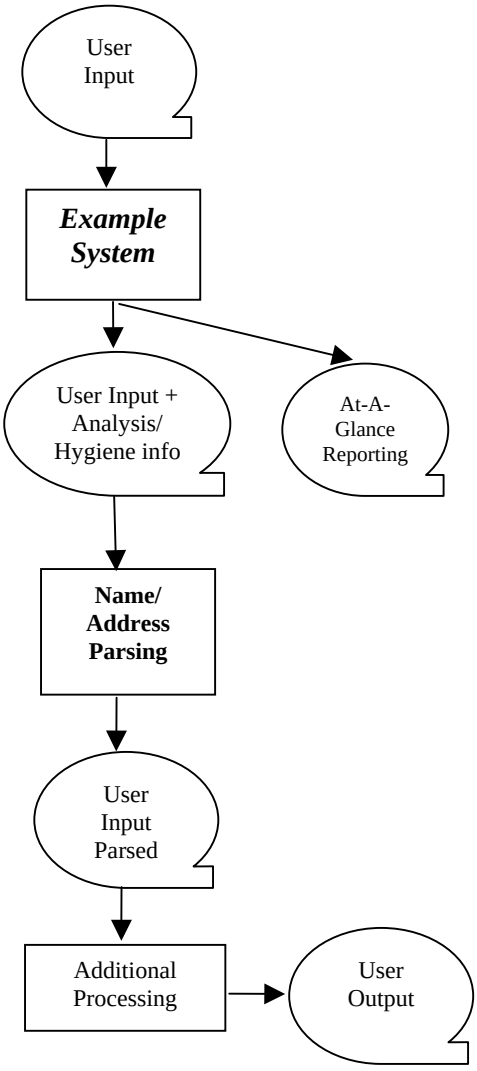

Figure 4: User Input, Unparsed

CONCLUSION

The increased necessity for quality name data in CDI processing, and the lack of resources and tools to implement supporting systems, can create a significant undertaking for any company. Our organization's focus in this area amplified the need for name quality in addition to delivery point validation resulting in great strides and success for many adopters. The challenges associated with the implementation of NameCheck have proven to be well worth the effort as rules and data from throughout the organization were leveraged providing users with a comprehensive data quality tool. NameCheck has proven to be an integral part of Acxiom's corporate-wide data quality system benefiting both the organization and marketplace with quantifiable evidence for assessment within the dimensions of data quality [1].

Future work focuses are to provide a more configurable version of NameCheck allowing users to enhance reference tables and reason codes application. Access to an environment supportive of development efforts utilizing large datasets will also allow future versions large reference table allocations. Furthermore, research into improved handling of the hygiened name functionality opportunities will be

explored.

REFERENCES

[1] Campbell, T. and Wilhoit, Z. "How's Your Data Quality? A Case Study in Corporate Data Quality Strategy." *Proceedings: International Conference on Information Quality*, MIT, 2003.

[2] Huang, Kuan-Tsae, Yang W. Lee and Richard Y. Wang, *Quality Information and Knowledge*. Prentice-Hall. Upper Saddle River, NJ. 1999. p. 59-90

[3] Sankey, Camie "Name Game" Bodo's Lair. 1996-2000. (http://www.bodo.com/jokes/jname.htm)