USING COMMERCIAL DATA INTEGRATION TECHNOLOGIES TO IMPROVE THE QUALITY OF ANONYMOUS ENTITY RESOLUTION IN THE PUBLIC SECTOR

(Practice-Oriented)
Methodologies

John Talburt
University of Arkansas at Little Rock
John.Talburt@acm.org

Charles Morgan
Acxiom Corporation
cdmorg@acxiom.com

Terry Talley
Acxiom Corporation
ttalle@acxiom.com

Ken Archer
Acxiom Corporation
karche@acxiom.com

Abstract: This paper describes a method for using new customer data integration and customer recognition technologies that have been developed in the private sector to solve the problem of anonymous entity resolution across multiple, non-shared data stores held in government agencies. In the method described, new commercial technologies that link records using anonymous tokens and encoded links allow indicative information across disparate data stores to participate in the entity resolution process without violating information-sharing policies. The benefit is that the resolution can be more complete and accurate through the broader inclusion of indicative data from many agencies and sources. The method also posits a trusted broker, an agent or agency that serves two primary functions. The first is to maintain the cross-reference table of encoded links for the participating agencies. The second is to translate the representation of the anonymous tokens presented by one agency into their corresponding representations in the other agency in those cases where information sharing is permissible and authorized. Although this paper describes the case where the entities are persons and locations, the same method is generally applicable to a broader range of entity types.

Key Words: Anonymous Entity Resolution, Data Quality, Customer Data Integration, Information Sharing, Hash Token, Equivalence Class Tokens

INTRODUCTION

Entity Resolution is the process in which records determined to represent the same real-world entity are successively located and merged [1]. In Anonymous Entity Resolution, each entity is represented by a "token" (a symbol or string of symbols that is a place holder for the entity) from which the identity of the entity cannot be inferred or otherwise derived, i.e., the entity remains anonymous. Anonymous entity resolution can help provide a solution to the problem in which there is a legitimate need to gather and analyze disparate information about an entity, but at the same time, the identity of entity must be cloaked in order to conform to privacy policies or legal regulation.

The attributes that are used to determine whether two entities are the same are called "indicative information." An important commercial application of Entity Resolution is Customer Recognition, where the entity in question is a customer, usually an individual or a business [5]. In recent years, businesses have realized that in a highly competitive environment that they must not only gain market share, but they must also retain and maximize the value of the customers they have. A company will have multiple interactions with the same customer at different times, locations, or lines of business. Each failure to connect these interactions is a lost opportunity to make them more profitable for the business and more satisfying for the customer. The collection of strategies around maximizing the value of these customer interactions is called "Customer Relationship Management," or CRM [7].

BACKGROUND

The Problem

Businesses routinely bring together customer data from disparate locations and lines of business into data warehouses. Commercial data warehousing is a mature, multi-billion dollar industry in the US, and technology companies have developed many sophisticated tools and products to facilitate their implementation and operation [3]. The ability to mine information afforded by the data warehouse concept has led to tremendous gains in productivity and efficiency for American businesses.

Government agencies have begun to realize that they could reap the same kinds of benefits [12]. However, the majority of public sector agencies are prevented from taking a data warehouse approach to data mining because of regulatory restrictions that limit their ability to consolidate or share information with other agencies.

Government agencies are finding themselves on the horns of a dilemma – "Be more efficient and effective in the use of data, but don't violate information sharing policies and regulations." The attack of September 11, 2001, and ensuing events have brought particular focus on this problem. On the one hand, there is increasing pressure to share information in order to "connect-the-dots" on terrorism [2], while on the other, there is concern about eroding our basic principles of freedom and rights to privacy [14]. The ability to assemble entity information across multiple sources and agencies could contribute to more complete and precise information, e.g., for "persons of interest." The problem is how to accomplish this while honoring all relevant information-sharing policies and regulations.

Even though the use of anonymous tokens can mitigate many of the issues surrounding information sharing, not all commercial methods and approaches to this problem are the same. Each will have different degrees of fitness (quality) with respect to several well recognized dimensions of data quality, including security, completeness, accuracy, consistency, and timeliness [15].

However before getting too deeply into the data quality issues, some background on anonymous entity resolution is in order. Token-based entity resolution systems fall into two broad classes, based on how the tokens are created:

- Hash Tokens, and
- Equivalence Class Tokens

Hash Tokens

The simplest method for associating a token with an entity is to use an algorithm to calculate or derive a value for the token from the primary indicative information for the entity. The derived value is called a "hash token." For example, if the indicative information for a customer were "Robert Doe, 123 Oak St.," then the underlying binary representation of this string of characters can be put through a series of rearrangements and numeric operations that might result in a value represented by string of characters "r7H5pK2." It is not difficult to design a hash algorithm in such a way that it is virtually impossible to reverse engineer (derive the indicative information from the token), thus the desired anonymity can be obtained from hash tokens. Some methods even go even farther to assure anonymity by applying standard encryption techniques to hash values [4].

The use of hash tokens for entity resolution has two drawbacks, hash collisions and lack of consistency. Hash collisions occur when the hash algorithm operating on two different arguments creates the same hash token, creating a many-to-one mapping from indicative information to the token representations. There are number of mitigations for hash collisions and for entity resolution; this does not present a major obstacle [8].

The more serious problem is related to consistency. Hash algorithms are notoriously sensitive to very small changes in the argument string. For example, even though "Robert Doe, 123 Oak St." and "Bob Doe, 123 Oak St." represent the same customer, most hash algorithms will produce very different hash values for each. In order to mitigate this problem, some systems introduce pre-processing routines to "standardize" the argument string before the actual hash algorithm is applied. These routines can range from removing extra white space and punctuation to more knowledge-driven transformations such as changing "Bob" to "Robert" [4]. However in real world, the indicative information for the same entity can change dramatically. For example, "Jane Doe, 123 Pine St." can marry John Smith and move to a new address, resulting in "Jane Smith, 345 Elm St." resulting in two valid, but very different indicative records for the same person. In cases like this, no amount of name and address standardization could enable these two records to produce the same hash token.

Equivalence Class Tokens

The only way to improve the consistency of token assignments for these kinds of situations is to use a knowledge base approach [9, 11]. As knowledge is acquired that indicative information for an entity has changed, the new representation is stored along with other valid representations in a list, called an "equivalence class." Each equivalence class is assigned an arbitrary, but unique, token value that is not derived from a particular representation of the entity.

Table 1 shows how both examples described earlier can easily be accommodated using an equivalence class approach. Note that neither of the tokens, xH45nT and y7Bw6, are derived from hashing the representations, but have simply been selected from some pre-defined list of unique token values and "assigned" to these representations.

	Token	Representation
Equivalence Class xH45nT	xH45nT	Jane Doe, 123 Pine St
	xH45nT	Jane Smith, 345 Elm St
	xH45nT	J S Smith, 345 Elm St
Equivalence Class y7Bw6	y7Bw6	Robert Doe, 123 Oak St
	y7Bw6	Bob Doe, 123 Oak St

Table 1. Two Equivalence Classes

If we consider all of the possible entity representations as the underlying set *S*, then the rule that "two representations are assigned the same token *if*, and only *if*, they represent the same entity" defines an equivalence relation on *S* that partitions *S* into equivalence classes, i.e., all representations associated with the same token. Equivalence classes, equivalence relations, partitions, and other concepts from abstract algebra are not only descriptive, but they also provide important new analysis tools for problems such as customer recognition related to data integration and entity resolution [13].

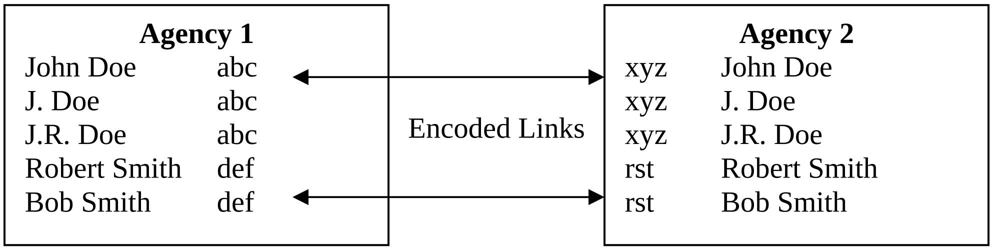
This same "if and only if" constraint described above also defines accuracy for entity resolution. Though closely related, consistency and accuracy are not the same. Just because variant representations of an entity are consistently assigned the same token it does not mean that the assignments are "correct." Without *a priori* knowledge, one cannot say whether "Jane Doe, 123 Pine St." and "Jane Smith, 345 Elm St." should correctly be assigned the same token or different tokens. Once these two representations are assigned tokens (correctly or incorrectly), consistency would dictate that their token assignments do not change.

As often happens, accuracy and consistency can be at odds. In practice, there may be latency in the source or sources of associative information. Consequently, the following scenario may ensue. The first representation of the entity is presented to the system and is assigned a new token value. Then the second representation is presented to the system, but there is no information available to connect it to the first, so the system assigns the second representation a different token. Later, new information arrives that connects these two representations causing the token assigned to the second representation to be discarded and replaced with the same token assigned to the first representation. In this case, system accuracy increases, but system consistency decreases.

Universality

Universality simply means "global consistency." Whereas local consistency says that all representations of the same entity will be assigned the same token within a given context (system or agency), universality or global consistency says that all representations of the same entity will be assigned the same token regardless of context. Although universality may seem like a good thing, it can work against security. Universal token assignment is in effect a "universal identifier" it provides back door that can be used to violate privacy and security protections. It would allow different agencies holding information on the same entity to easily collaborate independent of any third-party broker or mediator, thus defeating the original intent of the anonymous resolution.

A new technology that accommodates consistency within a system or agency, facilitates anonymous resolution, but does not require universality is "encoded links" [10]. In this method, each system or agency uses a locally consistent set of tokens, but the tokens for the same entity are not the same in different systems, i.e., not universal. The ability to localize tokens to a particular agency or domain is illustrated in Figure 1. In this sense, localization represents the inverse of universality.



Agency 1	
John Doe	abc
J. Doe	abc
J.R. Doe	abc
Robert Smith	def
Bob Smith	def

Agency 2	
xyz	John Doe
xyz	J. Doe
xyz	J.R. Doe
rst	Robert Smith
rst	Bob Smith

Figure 1. Encoded Links

Figure 1 shows that tokens are assigned consistently but differently within each agency. An encoded link is a cross-reference between the tokens representing the same entity in the two agencies. Resolution of information held by the two agencies for the same entity can still take place, provided that the encode link between the two tokens is known.

METHOD

The problem of entity resolution across non-shared data stores can be solved through a combination of commercially developed technologies for linking data with anonymous equivalence class tokens [9, 11] and encoded links [10] across different agencies. The remainder of this paper will focus on two types of entities: persons and locations, specifically postal addresses, even though the principles can just as easily be applied to other types of entities.

Overview

Figure 2 shows how the method can be described in terms of three actors that carry out the basic processes:

1) Participating Agency - stewards of the non-sharable data stores and the users of the entity resolution system
   a) Applies analytical processes and methods utilizing the anonymous tokens in its own data store, and when appropriate, links to entity information in the data stores of other participating agencies.
   b) Provides indicative information to the commercial processor for assigning anonymous tokens.
2) Trusted Broker - the agent or agency holding the cross-reference table of encoded links used by the participating agencies
   a) Receives each agency’s indicative information after anonymous tokens have been assigned by the commercial processor, encodes the links, and updates the cross-reference table.
   b) Translates encoded links among participating agencies when a legitimate request is made by one agency to access entity information held by other participating agencies.
3) Commercial Processor – provider of the anonymous token and encoded link technology
   a) Assigns anonymous tokens to the indicative information provided by each agency.
   b) Provides the trusted broker with the technology to encode links and maintain the cross-reference table.

For simplicity in Figure 2, only two participating agencies are shown.

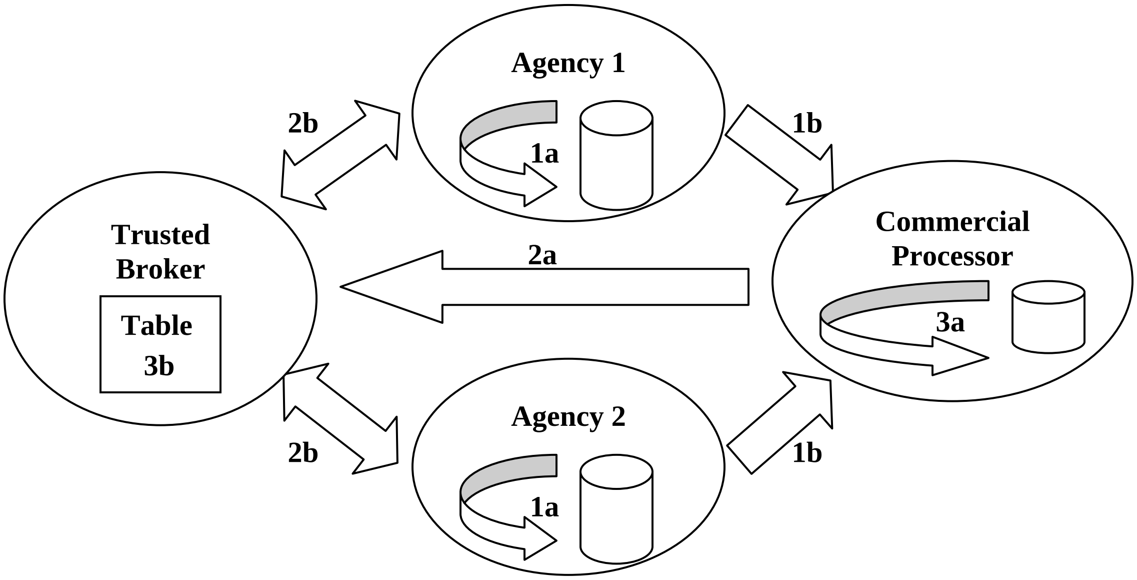


Figure 2. Context Diagram

Assigning Anonymous Tokens

A critical component of the method is the ability to assign anonymous, equivalence class tokens to each person and location. Each token is a unique combination of alphanumeric characters that is unrelated to the content of indicative information for the entity. Because it is not possible to reverse engineer indicative information from the value of the token; the token is anonymous. The purpose of the token is only to provide linkage among the items of information related to the same entity.

The assignment of an anonymous token is accomplished using a three-step process:

Step 1 Match the indicative information to a knowledge base of commercial, publicly available information. If a match is found, assign the token from the public knowledge base; otherwise, continue with Step 2.

Step 2 Match the indicative information to a registry of “agency-only” information, i.e., information found only in the stores of participating agencies. If a match is found, assign the token already in the registry; otherwise, continue with Step 3.

Step 3 Register the indicative information as agency-only, and assign a new token.

Example 1

Agency 1 provides two indicative records from its data store: Record 1 = “Mary Jones at 123 Oak, Anytown” and Record 2 = “Mary Johnston at 456 Elm, Anothertown.”

When Record 1 is processed, a person and address match are found in the public knowledge base in Step 1, and the person token “**pA1B2C**” and address token “**aX5Y6Z**” is assigned to the record.

When Record 2 is processed, a person and address match are also found to the public knowledge base in Step 1, and furthermore it is for the same person. Therefore the same person token "**pA1B2C**" is assigned to this record. The address is also found in the public knowledge base, but it is a different address so a different address token "**aJ4K5L**" is assigned to the record.

This example illustrates the case in which the same person has changed both name and address. Consequently, the token assigned to the person is the same for both records (same person entity), whereas the tokens for the address are different (different address entity).

Example 2

Agency 1 provides Record 1 = "Sam Brown at 29 Pine, Anytown" and Agency 2 provides Record 2 = "Sam Brown at 29 Pine, Anytown."

When Record 1 is processed, a match is found for the address in the public knowledge base and is assigned address token "**aT7Y4N**," but no person match is found. Therefore, processing continues to Step 2.

At Step 2, no match is found for the person in the agency-only registry. Therefore, a new person token is created, "**pQ3W4E,**" and assigned to record. In addition, the indicative information and token are added to the agency-specific registry.

When Record 2 from Agency 2 is processed through Step 1, the same address is again found in the public knowledge base, and the same address token "**aT7Y4N**" is assigned. Similarly, the person is still not found in the public knowledge base.

However at Step 2, a match is found for this person in the agency-only registry, and the previously created registry token "**pQ3W4E**" is assigned.

This example illustrates another advantage afforded by the method. By first matching to the public knowledge base, many persons and addresses can be resolved immediately utilizing the large amount of public information available. However if a record does not match the public knowledge base, the agency-unique information is not added back to the public knowledge base, but is maintained separately in the agency-only registry. In this way, the information remains available to match information from other agencies, but is maintained separately from publicly available data. This allows the consistent assignment of tokens across agencies without commingling agency unique information with public information.

Real-Time Recognition

Another advantage of equivalence class token assignment is that resolution is primarily a matter of table look-up operations. When deployed on a high-capacity, high-performance platform such as a grid-based system, all of the operations described above can be carried out in real time for large volumes of data in many agencies.

Example 3

An agent in the field from Agency 1 enters an entity record on a wireless handheld device that queries the Agency 1 data store. At the same time, the Agency 1 system recognizing that the policy governing this request and entity type also allows it to utilize any information for the entity that might be held by Agency 2, if available. The system then automatically forwards the request for information on this entity from Agency 2, along with the Agency 1 tokens, and appropriate authorization information to the Trusted Broker. The Trusted Broker authenticates the request from Agency 1, translates Agency 1 tokens into Agency 2 tokens, and forwards the translated request to Agency 2. Agency 2 results are routed back to Agency 1, and the results from the analysis of the information held by both Agency 1 and 2 for the entity are then returned the field agent in real time.

Encoded Links

In addition to anonymous tokens, a second technology is employed to explicitly control the sharing of information. In order to maintain integrity and consistency of the tokens, the commercial processor must maintain and assign only one set of unique tokens in the process described above. If these tokens were to be returned directly to the agencies, the result would be the virtual federation of all entity information across the participating agencies based on a universally assigned token. To prevent this from happening, the commercial processor does not return the assigned tokens to the agency providing the indicative information. Instead, the tokens are given to the trusted broker, where they are encoded into alternate representations unique to each agency.

In addition, the trusted broker must also be able to maintain the proper mappings between the alternate representations used by each participating agency. Thus, the correspondence between the representation of a token in one agency and the representation of the token for the same entity in another agency comprises an encoded link. The mappings or encoding are only available to the trusted broker. In this scheme, only the trusted broker has the ability to translate the representation of a token belonging to one agency into the representation used by another agency, and thus controls the ability to request, and ultimately, share information across participating agencies.

Example 4

Continuing with Examples 1 and 2 above, if these data were processed and returned to the trusted broker as shown, then the result would be the creation of two tables as illustrated in Table 2.

Native	**Agency 1**	**Agency 2**
pA1B2C	**p65YG4**	**p3TGEG**
pQ3W4E	**pL3H4H**	**pY94BN**

Table 2. Person Cross-Reference

This example illustrates how the trusted broker must not only maintain a list of person and address tokens that occur in the participating agencies, but it must also create alternate representations for each agency and maintain their correspondences. Again, all tokens are anonymous and bear no derived relationship to the underlying indicative information they represent. The only requirements are that the alternate representations must be consistent within each agency and the mapping from the native representation to the agency representation must be one-to-one.

Example 4

Continuing with Example 2 and Example 3 above, Agency 1 holds information in its data store on the entity "Sam Brown at 29 Pine, Anytown," and from Table 2, that entity is associated with the token "**pL3H4H**" in the Agency 1 data store.

If Agency 1 has a legitimate reason to gather information about this entity from Agency 2, then Agency 1 submits this request along with appropriate authentication and token to the trusted broker along with the token "**pL3H4H**."

The trusted broker then validates the requests and translates Agency 1's token "**pL3H4H**" into Agency 2's representation of "**pY94BN**."

Agency 2 can then interrogate its data store to see if it holds any information for this entity using its internal representation of "**pY94BN**."

This example illustrates how encoded links provide protection from direct sharing of information based on tokens. Because tokens are consistent within an agency, bringing together information related to an entity within the same agency can easily be done through a straight forward, token-based query. However, the encoding of the links prevents this same operation from working across agencies. Each agency can query on its representation of a token, but it must rely on the trusted broker to get information from other agencies.

CONCLUSION

Using the new commercial technologies of equivalence class token assignment and encoded links, it is possible to construct a method for anonymous entity resolution across government agencies while still supporting policies and regulations related to inter-agency information sharing and privacy protection. Moreover through the utilization of this method, improvements to data quality can be realized over several dimensions.

Security

The proposed method improves data security in two ways. The use of arbitrarily assigned equivalence class tokens insures that anonymous representation of the entities. The theft or interception of transmitted tokens provides no information about the entities they represent.

Secondly, assigning tokens that are only consistent within a local agency creates a barrier to direct, token-based information sharing between agencies. At the same time, token-based information sharing between agencies for authorized transactions is actually made easier through the use of encoded links that are provided by the trusted broker.

Completeness

By simply providing a reliable mechanism that allows multiple agencies to share information appropriately and securely, a much larger, though virtual, store of information becomes available. In the case of entity resolution, more is better. By consolidating information from multiple sources, a more complete picture of the entity can be formed.

Accuracy

As with most data mining applications, greater completeness results in greater accuracy. Although multiple sources may result in some conflicts that must be resolved, analysis and decision-making are generally better with more information available than less.

The use of equivalence class tokens also contributes to more accurate entity resolution by accommodating a much broader range of indicative information than is possible using hashing methods. Furthermore, by maintaining the lists of indicative variants, it is much easier to re-group records once additional information does become available.

For example, records with the indicative information “Robert M Doe, 123 Main St” and “Bob Doe, 123 Main St” might initially be grouped together. However, the broader inclusion of information may show an apparent conflict in that the date of birth reported for “Robert” is 25 years earlier than the date of birth reported for “Bob.” Subsequent information may then shows that “Robert M Doe” is really “Robert Doe, Sr” and “Bob Doe” is really “Robert Doe, Jr” living in the same household. If this were the case, the original equivalence class can easily be split to assure that records indicating “Robert M” are more accurately assigned a different token to represent “Robert Doe, Sr.” This level of discrimination in token assignment afforded by equivalence class tables is very difficult to accomplish using hash algorithms.

Consistency

The application of equivalence class token assignment produces more consistent entity representation by removing the dependency on the format of the indicative data from the token value assignment.

Timeliness

Because the equivalence class token assignment is data (table) intensive, it can be slower than hash token assignments. However, state-of-the-art high performance computing (HPC) and high throughput computing (HPT) can provide substantial mitigation. For example, there is a very large commercial implementation of a name and address-based entity resolution currently running in grid computing environment [6]. The resolution service requires just over 200 grid nodes that manage about 1.5 billion rows of equivalence class tables, and can resolve as many as 700 million inputs per hour.

REFERENCES

[1] Benjelloun, O., Garcia-Molina, H., Su, Q., Widom, J. “Swoosh: A Generic Approach to Entity Resolution.” Technical Report, Stanford InfoLab, March 3, 2005.

[2] Bush, G.W. “Strengthening the Sharing of Terrorism Information To Protect Americans.” Executive Order 13356, http://www.fas.org/irp/offdocs/eo/eo-13356.htm, August 27, 2004.

[3] Eckerson, W.W. “Data Warehousing in the 21st Century.” *What Works, 9.* May 2000.

[4] Frederich, A. “IBM DB2 Anonymous Resolution: Knowledge Discovery Without Knowledge Disclosure.” IBM White paper. May 2005.

[5] Hughes, A.M. “Building Customer Loyalty by Recognition.” Database Marketing Institute, http://www.local6.com/news/4643968/detail.html, June 23, 2005.

[6] Jordan, W., “Acxiom Makes Grid Reality,” *Database Trends and Applications*, April 2004.

[7] Lee, D. *The Customer Relationship Management Survival Guide*. HYM Press. 2000.

[8] McKenzie, B.J., Harries, R., and Bell, T., “Selecting a Hashing Algorithm,” *Software- Practice & Experience*, 20(2), February 1990, pp. 209-224.

[9] Morgan, C.D., McLaughlin, G.L., Fogata, M.G., Baker, J.L., Cook, J.E., Mooney, J.E., Roland, D.B., Talburt, J.R. "Method and System for the Creation, Enhancement, and Update of Remote Data Using Persistent Keys." US Patent 6,073,140, June 6, 2000.

[10] Morgan, C.D., Talburt, J.R., Harvey, S., Talley, T., Anderson, W., Welch, S.K., White, C.S. “Data Linking System and Method Using Encoded Links.” US Patent Application Serial 10/217,059, August 12, 2002, European Patent Office Application Number 03254567.5, July 21, 2003.

[11] Morgan, C.D., Talley, T., Talburt, J.R., Bussell, C., Kooshesh, A., Anderson, W., Johnston, K., Farmer, R., Hashemi, R., Dobrovich, M., Baxter, R., Ward, M.K., Ratliff; J.K. "Data Linking System and Method Using Tokens." US Patent 6,523,041, February 18, 2003.

[12] Sinrod, E.J. “What’s Up with Government Data Mining?” *USA Today*, June 9, 2004.

[13] Talburt, J., Kuo, E., Wang, R., Hess, K. "An Algebraic Approach to Data Quality Metrics for Customer Recognition." *Proceeding of the 9th International Conference on Information Quality (ICIQ-2004),* MIT, Cambridge, Massachusetts, November 5-7, 2004, pp. 234-247.

[14] Venton, D. “Government Data Mining Raises Privacy Concerns,” *Computerworld*, January 17, 2003.

[15] Wang, R.Y., and Strong, D.M., “Beyond Accuracy: What Data Quality Means to Data Consumers,” *Journal of Management Information Systems* (JMIS), 12(4), 1996, pp. 5-34.