KNOWLEDGE-DRIVEN IDENTITY RESOLUTION FOR LONGITUDINAL EDUCATION DATA

A Dissertation Submitted
to the Graduate School
University of Arkansas at Little Rock

in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Applied Science

in the Department of Applied Science
of the Donaghey College of Engineering and Information Technology

May 2010

Greg Holland

B.S., University of Arkansas at Little Rock, 1998
M.S., University of Arkansas at Little Rock, 2007

This dissertation, "Knowledge-Driven Identity Resolution for Longitudinal Education Data", by Greg Holland, is approved by:

Dissertation Advisor:

John R. Talburt
Professor of Information Science

Dissertation Committee:

Daniel Berleant
Associate Professor of Information Science

M. Keith Hudson
Professor of Applied Science

Elizabeth M. Pierce
Associate Professor of Information Science

Ningning Wu
Associate Professor of Information Science

Program Coordinator:

Abhijit Bhattacharyya
Associate Professor of Applied Science

Graduate Dean:

Patrick J. Pellicane
Professor of Construction Management

Fair Use

Duplication

I authorize the Head of Interlibrary Loan or the Head of Archives at the Ottenheimer Library at the University of Arkansas at Little Rock to arrange for duplication of this dissertation for educational or scholarly purposes when so requested by a library user. The duplication will be at the user's expense.

Signature_____

KNOWLEDGE-DRIVEN IDENTITY RESOLUTION FOR LONGITUDINAL EDUCATION DATA, by Greg Holland, May 2010

ABSTRACT

Data sets containing information for an overlapping group of real-world identities present a very high likelihood that the identifying attributes and attribute values for these identities may be inconsistent between the data sets. Differences in the types of identifying attributes or attribute values inhibit proper record linkage and identity resolution. Traditional approaches to record linkage are commonly utilized however the results from these approaches do not demonstrate the highest possible levels of confidence and utility. Syntax, semantics, and temporal aspects of data sets should be understood and incorporated into the methodology of heterogeneous data set integration. Domain-specific expertise is a key component of methodology development. The goal of this research is to determine a course of action which will facilitate knowledge-driven identity-resolved longitudinal data studies with optimal record linkage for data sets containing varying identifying attributes and attribute values obtained through various collection methods over a number of years. The proposed identity resolution methodology will be demonstrated with four years of actual education data for students within Arkansas Department of Education data sets. This research will facilitate a FERPA-compliant plan for resolving the representations of real-world identities across multiple longitudinal education data sets, allowing for record linkage of statewide education data and increasing the capability of various state agencies to coordinate future research efforts for education data.

To my wife, Laura

&

My parents, Carrell and Vesta

Acknowledgements

I would like to thank my advisor, Dr. John R. Talburt, for his continued guidance in my doctoral research and the years of valued discussion and instruction since our earliest work together in 1998. My gratefulness also extends to my Ph.D. committee members and their provision of constructive suggestions, encouragement, and ideas for my research.

I would like to thank my wife, Laura, for her continuous support, love, and encouragement. I also would like to thank my mother and father for their support in all my academic endeavors throughout my life. They set the example for the importance of education in our family.

I have benefited greatly from colleagues at the two organizations with which I have been employed over the past decade. The challenges and collaborations provided by these individuals have driven me to overcome obstacles, research possible approaches, and find innovative solutions. Special thanks are extended to fellow graduate student and co-worker Neal Gibson from the Arkansas Department of Education.

It is my privilege to be among the first to pursue advanced degrees in the emerging field of Information Quality, and I am grateful to those visionaries who have provided many years of research as the foundation for the establishment and future of this discipline through the Massachusetts Institute of Technology programs and conferences on Information Quality.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Problem Definition

Data sets containing information for an overlapping group of real-world entities (particularly when those entities are people) present a very high likelihood that the identifying attributes and attribute values for these entities are not completely consistent between the data sets. Though it is understood that the same entities may be represented in several data sets, there may be differences in the types of identifying attributes or attribute values provided. These inconsistent entity references are a pervasive information quality problem in all heterogeneous data set integration efforts. Differences in attribute values occur for syntactical, semantic, and temporal reasons. Entity resolution or semantic integration is the resolution of semantic conflicts that disable a one to one mapping between concepts or terms (Bijlsma, Koolwaaij, Schoneveld, Nuijten, & Schaafsma, 2002). Integration methods which utilize a knowledge-driven approach to identity resolution provide significant advantages over the traditional methods for record linkage. Real-world applications are continuing to increase in number, as organizations improve computing capabilities and recognize the need for a more comprehensive view of their data assets.

The inability to compare different identifier attributes is the primary problem in attempting to link records between these types of data sets. A secondary problem is still present when attributes are common to both sets, such as the last name, because the data may still contain variations in spelling, format, or type of the attribute values referencing the same entity. It is necessary to resolve the representations of the same real-world

entities, in spite of the differences in the data sets. The absence of identifiers for the underlying entities often results in a database which contains multiple references to the same entity (Bhattacharya & Getoor, 2007).

The variations in the identifying elements are only part of the obstacles to proper record linkage between data sets. Another obstacle is the changing nature of the real-world entities (people) themselves. Name changes occur with marriage, divorce, adoption, and other circumstances which will hinder the proper identification of entities even when the same name attributes are utilized as identifiers. In the event that no changes to the real-world entities have occurred, and in the event that the same identifying elements are utilized between two data sets, there may still be obstacles to proper record linkage when data collection or compilation errors have occurred. Misspellings, character transpositions, and other errors in one or more attribute values introduce additional difficulty for identity resolution between two data sets. Databases may contain duplicate records concerning the same real-world entity because of data entry errors, because of un-standardized abbreviations, or because of differences in the detailed schemas of records from multiple databases, among other reasons (Monge & Elkan, 1997).

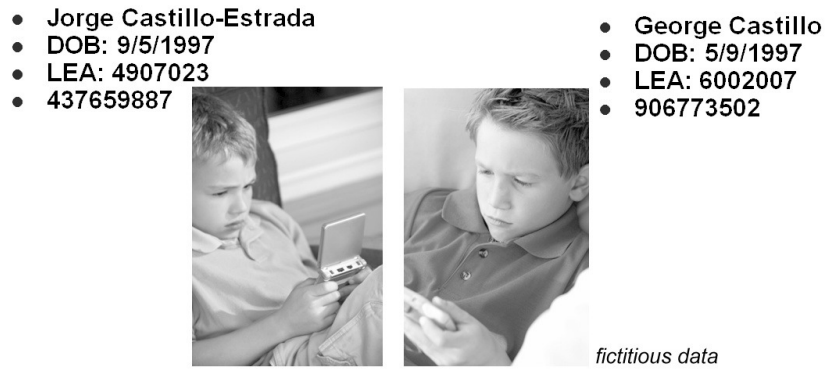

Figure 1. Example of attribute value changes common to identity resolution processes.

The key to effective identity resolution is emulating an intelligent user's ability to determine a match based on a variety of factors, overcoming spelling, phonetic, and other errors and omissions in the data while offering the speed and scale to perform high-volume searches quickly against very large databases (Informatica, 2008). This research will describe the methodologies which are proposed to achieve both of the organizational goals mentioned above: to increase the quantity of identity resolutions and to increase the speed at which these resolutions can be obtained.

The Arkansas Department of Education has struggled in systematically obtaining student identity resolutions despite the subject matter expertise of its leading researchers. Traditional record linkage efforts have been effective, but have not displayed the capabilities that stakeholders would like to see in both the quantity of resolutions and speed at which those resolutions can be obtained (Tachinaba & Garcia-Molina, 2009). One obstacle to current identity resolution efforts involves the heterogeneity of the data sets in question. In one education data set the first name, last name, and Social Security Number may be the only identifying attributes. In another data set the last name, date of

birth, and a local identification number may be the only identifying attributes. Because each of these data sets may be structured according to the requirements of particular data collection or management systems, it is understandable that the identity attributes are different and may not be adjustable.

To demonstrate the effectiveness of the methodologies proposed in this research, actual data will be utilized for identity resolution within the Arkansas Department of Education. Multiple heterogeneous data sets will be utilized, which have been collected and maintained by the Arkansas Department of Education for students in kindergarten through twelfth grade (K-12). These multiple data sets will be longitudinal in nature, consisting of the same real-world entities represented in each data set annually, if not more frequently. The data presents a span of time for each entity, with the understanding that entities should have a continuous history of representation in the time-sequential data sets. The varying identifying elements utilized in these multiple data sets present the entity resolution problem, which must be resolved in order to accurately identify the real-world entity representations and understand the historical data in a longitudinal format.

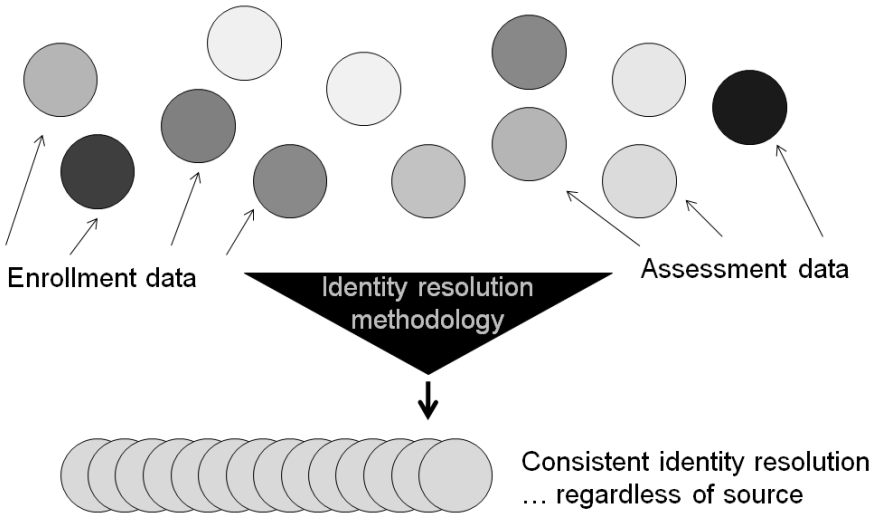

Figure 2. Illustration of heterogeneous data sets obtaining consistent identity resolutions.

The goal of this research is to determine a course of action which will facilitate identity-resolved longitudinal education data studies with optimal record linkage for data sets containing varying identifying attributes and attribute values, obtained through various collection methods over a number of years. This research will implement a methodology for resolving the representations of real-world entities (students) across multiple longitudinal education data sets which do not utilize a consistent set of identifying attributes, allowing for student record linkage of statewide education data and increasing the capability of various state agencies to coordinate research efforts for education data from the past decade.

1.2 Literature Review

Record linkage has been recently defined as "simply the bringing together of information from two records that are believed to relate to the same entity - for example, the same individual, the same family, or the same business" (Herzog, Scheuren, &

Winkler, 2007). The basic concepts behind record linkage are thousands of years old. We use the term "recorded history" specifically because of the invention of writing and the introduction of records of people and accounts of the ancient world. The first reconciliation of these records in order to increase the understanding of the subject would have been the first "record linkage", though we do not know when that might have occurred. Nevertheless, throughout history records have been compiled, reconciled, and summarized utilizing some form of record linkage.

Record linkage as a modern practice is attributed first to Halbert L. Dunn. In 1946, Dunn published a paper entitled "Record Linkage" as a result of his work as Chief of the National Office of Vital Statistics. Dunn used the analogy of a "Book of Life" for each person in the world, with record linkage as the process needed to assemble the pages of each book (Dunn, 1946). Dunn was primarily concerned with accurate recording of births and deaths in order to save the United States millions of dollars in maintaining active files for the deceased and also to aid the insurance industry in death benefit payments. Though this process was still manual, the seeds of the electronic methods on the horizon were planted by Dunn.

The practice of record linkage was discussed widely by H. B. Newcombe in relation to his work involving vital records in the 1950s. In obvious reference to Dunn's "Book of Life" example, Newcombe (1959) noted "the various facts concerning an individual which in any modern society are recorded routinely would, if brought together, form an extensively documented history of his 1ife. In theory at least, an understanding might be derived from such collective histories concerning many of the factors which operate to influence the welfare of human populations, factors about which we are at

present almost entirely in ignorance." Newcombe's work through the 1960s defined the logical methods utilized at the time. "The two principal steps in any linking operation, namely, those of searching out the potentially linkable pairs of records for detailed comparison and of deciding whether or not a given pair is correctly matched, are commonplace in almost any operation by which a file is kept up-to-date" (Newcombe, 1967).

Newcombe's concepts are still in place today, though the methodologies used to identify potentially linkable pairs involve several different approaches. Common record linkage methods include: exact attribute value matching, merging, edit distance, Soundex, similarity matching, common string frequencies, and clustering ("Record linkage", 2010). Organizations apply one or more of these methods to existing data sets in order to determine multiple instances of real-world entities within the data. Each of these approaches has been studied in much detail and each has strengths and weaknesses. Some of those limitations will be discussed in the next section of this research.

In the earliest decades of research, only the term "record linkage" was utilized in the literature, however, a particular aspect of record linkage is identity resolution or entity resolution. The aspects associated with identity resolution are specific to the information age of the most recent decades and the proliferation of computing capabilities for all organizations, not just organizations in the technology industry. Identity resolution is defined as an operational intelligence process, typically powered by an identity resolution engine or middleware stack, whereby organizations can connect disparate data sources with a view to understanding possible identity matches and non-obvious relationships across multiple data silos ("Identity resolution", 2010). While this definition may be

accurate, the underlying concepts of record linkage still apply. Identity resolution still involves the seeking out of potentially linkable pairs, and implements a decision about the accuracy of those potential linkages.

The core concept for identity resolution (entity resolution) is that the records are representative of some real-world entity. Usually, the entities being resolved are associated with people, though it is also possible to resolve the identities of organizations, locations, documents, or products. The goal of identity resolution is to accurately and consistently identify references to the real-world entity. This process involves the integration of information related to these entities (Talburt, Wu, Pierce, & Hashemi, 2007). Two specific possibilities arise whenever potentially linkable pairs are identified: an incorrect linkage may occur, or the decision to leave the identities separate may be incorrect (Statistical Society of Canada [SSC], 2008). A third possibility is that a possible link has been identified, but resolution cannot be determined at the time (Fellegi & Sunter, 1969). Researchers from Columbia University and AT&T Labs have stated, "for reasons of correctness and efficiency, we require no false dismissals and few false positives respectively" (Gravano, et al., 2001). It is not clear that this opinion is universally accepted, and a number of limitations to existing methods are known.

1.3 Limitations of Existing Methods

As mentioned in the prior section, common entity resolution and record linkage methods include: exact matching, merging, edit distance, Soundex, similarity matching, common string frequencies, and clustering. Records in the two data sets are standardized, compared, and one or more (often proprietary) methodologies is applied in order to identify the same entities represented in both data sets. This process involves "exploiting

cues from several sources, including, frequently occurring words within an element, partial sequential relationships amongst elements, length distribution of elements, and external databases of relationship amongst symbols" (Borkar, Deshmukh, & Sarawagi, 2001). The limitations of this process are numerous. As Newcombe (1967) explained, "the problem is that of enabling the machine to apply in numerical form the rules of judgment by which a human clerk would decide whether or not a pair of records relates to the same person when some of the identifying information agrees and some disagrees." As is usually the case with a mathematical model, the model does not, in every respect, faithfully represent the real world that it is intended to describe (Tepping, 1968).

Although entity resolution often incorporates some phonetic logic or nickname tables in order to identify names which sound alike or are known aliases (such as 'James' and 'Jimmy'), these rules are generalities only (Varol, 2009). When an organization requires identity resolution for data sets containing the same real-world entities, the organization should supply as much real-world knowledge as possible to the resolution process. For example it is common knowledge that 'James' and 'Jimmy' may be aliases for the same individual, however, a particular organization may be aware of cases where a particular 'James' also goes by the name 'Patrick' in more formal situations. Perhaps the person's full name is 'Patrick James Smith'. The limitations of phonetic logic and nickname tables in identity resolution are quickly apparent when particular individuals are known to be represented in an organization's data sets without incorporating the organization's own reference data (Christen, 2006). As a result confidence in the resolution process may be limited. Members of the Stanford Entity Resolution Framework (SERF) have noted, "even though Entity Resolution is a central problem in

information integration, and even though confidences are often an integral part of resolution, relatively little is known about how to efficiently deal with confidences" (Menestrina, Benjelloun, & Garcia-Molina, 2006).

Limitations arise whenever the identifying attributes are inconsistent between heterogeneous data sets. If one data set uses first name, last name, and date of birth, while another data set uses a local identification number, last name, and Social Security Number, the only attribute in common is last name. While it is possible that a portion of the real-world entities in both sets possess unique last names, common last names such as 'Smith', 'Jones', etc., will limit accurate identity resolution among the data sets. A third source of information is required; a reference data set which includes all five attributes would be particularly beneficial in this identity resolution effort. If no reference data is available, the identity resolution efforts will fail under these circumstances. Newcombe was resigned to this fact even late in his career. "Accuracy was strongly dependent on the amount of personal identifying information available on the records being linked" (Newcombe, et al., 1983).

Identity resolution efforts may include a special case which is both limiting and beneficial. Whenever all of the entities represented in one data set are known (assumed, believed) to also be present in a second data set, a "closed system" has been identified. This situation occurs often in longitudinal data whenever a set of individuals are represented over time in multiple data sets. Identity resolution is often handled at the individual level only; a particular entity is presented to the system, and the system attempts to resolve to the best available identity from one or more candidates. The results of this resolution are independent of the prior or successive resolutions attempted by the

system. Several authors have pointed out that match decisions should not be made independently for each candidate pair (Singla & Domingos, 2006).

A closed system allows the resolution process for the complete data set to be utilized in deciding the identity of each entity. Whenever each entity is known (or believed) to be present in the second data set, a process-of-elimination methodology can enhance the identity resolution efforts. Generally speaking, identity resolution and record linkage methods do not often utilize any prior knowledge about the entities which are linked between two data sets in the majority of today's methodologies. Collective entity resolution improves performance over independent pair-wise resolution (Bhattacharya & Getoor, 2006). It is the collective entity resolution approach that facilitates the knowledge-driven identity resolution proposed in this research.

The purpose of identity resolution is to emulate the decision-making process of a knowledgeable person who is tasked with determining whether the identities of two records in fact refer to the same real-world entity. The steps associated with this process are then automated in the hopes that the resulting identity-resolution system is significantly faster than the manual process, while maintaining the accuracy of the human knowledge-driven examples.

CHAPTER 2

PLANNING IDENTITY RESOLUTION

2.1 Proposed Approach

One goal of this research is to overcome the limitations identified in the prior section through the use of an organization-specific reference data set, taking advantage of the closed system and longitudinal aspects in addition to the traditional "tried-and-true' approaches to identity resolution. This research will not only resolve the identities of longitudinal education data, but it will also demonstrate the methodology by which other organizations will be able to apply similar approaches. These other organizations may include agencies in the same state or departments of education in other states. The resulting reference datasets will form the foundation of an on-going identity resolution system which will be utilized by the Arkansas Department of Education (ADE) internally and in coordination with other state agencies.

As a research analyst and project manager at ADE, access has been granted to longitudinal education data in various data sets. These data sets represent actual education data for the state of Arkansas for students over a period of four years.

Table 1 Data sets utilized in this research.

Data Set	Quantity
Student Enrollment 2005-2006	590,806
Student Enrollment 2006-2007	584,098
Student Enrollment 2007-2008	588,279
Student Enrollment* 2008-2009	463,405
ACT (College Board) FY 2007	48,258
ACT (College Board) FY 2008	50,376
ACT (College Board) FY 2009	56,611
Explore 2008	25,119
Explore 2009	24,447
Plan 2008	25,442
Plan 2009	26,016
TOTAL	2,434,599

* database changes in 2009 impacted quantity

When a third-party vendor's multi-year enrollment records are also included in the implementation, the total volume of records requiring some form of identity resolution in the future of this research exceeds 10,000,000. Additionally, some earlier data may be available in limited quantities. The resolved entities from these data sets will be utilized extensively by ADE and other state agencies as reference data in the proposed knowledge-driven identity resolution system for longitudinal data.

2.2 Key Considerations

The key considerations and constraints of this type of data research are related to the identifier attributes. The factors proposed for this identification of candidate indicator attributes are:

- The attribute must be present in both the input and reference data sets.
- Indicative attributes of interest when resolving identity will have a high percentage of distinct values, as well as a high percentage of unique values.
- Combinations of indicative attributes selected by the second factor will increase the percentage of distinct and unique values.

Explanation of the first factor: In order to be of value to the identity-resolution effort, any attribute to be utilized by the system must be present in both the data set containing entities to be resolved and the data set which is being referenced. If an attribute is present in only one of the data sets, it cannot be utilized when resolving identities. For example, a date of birth may be very valuable to resolve the identity of students, however, a data set which does not contain date of birth cannot benefit from date of birth knowledge.

Explanation of the second factor: Similar to the process to determine a primary key for a data set, attributes which are unique to each entity provide the most value to identity-resolution efforts. Unlike the primary key identification process, however, there is still value in determining which attributes provide a large percentage of unique values for the data set. Even when an attribute value is not unique, the number of candidate entities is lowered by the utilization of the attribute.

Explanation of the third factor: Though there is no guarantee that a date of birth is unique to a particular student, and though it is understood that multiple students attend the same school, the combination of the school and date of birth may provide a unique value which can be utilized to identify a particular student. Neither attribute is necessarily unique for any student if utilized separately.

2.3 Utility and Confidence

Because heterogeneous data sets contain varying attributes, the greatest amount of utility will result from resolution rules which require only a single attribute. However, the confidence when utilizing only a single attribute must be lower than the confidence of utilizing two or three attributes to resolve identity.

For example, a student may be resolved using only the Social Security Number value, if the reference data indicates that only one student has ever been represented by the Social Security Number value. However, there is no guarantee that another student may not present the same Social Security Number, either as a result of a typographical error, an intentional forgery, or some other cause. Though the reference data would provide statistical confidence that the identity was correct, additional confidence would be justified if the student's last name also matched the reference data. In the event of the typographic error on the Social Security Number, the last name's low likelihood of match to the incorrect student would prevent an incorrect identification or consolidation. First and last names may be of utility to identification efforts, however care should be taken to remove more frequently used name combinations from any list of potential matches based upon first and last name.

Table 2 Most common first and last name combinations

Student Count	First Name	Last Name
64	JOSHUA	SMITH
56	ASHLEY	SMITH
52	JESSICA	SMITH
48	JUSTIN	SMITH
37	ASHLEY	JONES
31	JUSTIN	WILLIAMS
30	JESSICA	JOHNSON
27	JOSHUA	BROWN

Added confidence is also warranted whenever additional attributes are utilized in determining the identity of the student. Though a particular Social Security Number and last name combination may be unique in the reference data, the utilization of the first name, date of birth, and other indicative attribute values would increase the confidence that the identification was correct.

Although the confidence increases through the use of additional attributes, the utility of these combinations will decrease. As mentioned earlier heterogeneous data sets do not contain the same attributes, and a strict requirement to include Social Security Number, last name, first name, date of birth, and other indicative attributes would be impossible whenever a data set does not contain one or more of these attributes. Utility for a particular high-confidence combination would be zero if one or more of the required attributes was unavailable. It becomes necessary to remove one or more of the attributes from the resolution methodology for that particular data set in order to obtain any utility of the data, however the removal of those attributes would decrease the confidence in the resolution results.

In summary, as the number of attributes included in identity-resolution methodologies increases, the confidence increases while the utilization decreases. Conversely, as the number of attributes included in identity-resolution methodologies decreases, the confidence decreases while the utilization increases. It is recommended that a formula be utilized or developed to provide metrics for both the utilization and confidence of identity-resolution processes.

2.4 Identity-Resolution Scenarios

Generalized scenarios that require split or consolidation are demonstrated in the tables below. In these examples, Entity 1 should be associated with ID1, and Entity 2 should be associated with ID2.

Table 3 Over-consolidation

	Data Set 1	Data Set 2	Conclusion
Entity 1	ID1	ID1	1.A. Correct
Entity 2	ID2	ID1	1.B. Incorrect over-consolidation

Table 4 Under-consolidation

	Test 1	Test 2	Conclusion
Entity 1	ID1	ID1	2.A. Correct
Entity 2	ID2	ID3	2.B. Incorrect under-consolidation

Though over-consolidation and under-consolidation represent inaccuracies in resolution, the number of over-consolidations should be minimized. It is much easier to combine the data associated with two entities once it is determined that they are identical than it is to separate the data associated with two entities which have been combined

incorrectly in the past. A comprehensive record of the sources and origins for each piece of information related to the two entities would be required in order to correctly undo the damage of an inappropriate consolidation.

An automated system providing identity-resolution based upon trusted reference data is being proposed in this research. The purpose of the system is to resolved identity in multiple longitudinal data sets, facilitating individual-level research. In the future the identity-resolution system will act as a trusted broker of identity information, allowing multiple state agencies to share information about the same entities without violating existing privacy laws.

2.5 Impact of Proposal

The Federal Educational Rights and Privacy Act of 1974 (FERPA) prohibits individually-identifying information from being shared between agencies. The indicative attributes of the student data in this research is protected by FERPA, and cannot be legally shared with other state agencies, such as the Department of Higher Education or the Department of Workforce Services. Several research proposals and opportunities are hindered by the inability to correctly identify individuals across agencies without violating FERPA.

A trusted-broker system utilizing a knowledgebase of reference data for Department of Education data could allow these agencies to resolve identities and associate those identities with non-personally-identifiable values, providing linkage without revealing the attribute values which are prohibited by law. Beyond the particular agency featured in this document, the methodologies presented in this research could be utilized by other agencies and other states, or by any organizations wishing to protect

individuals through the limited utilization of personally-identifiable attributes. Laws governing and limiting the use of sensitive data attributes are expected to increase in the future. The identity-resolved knowledge base of the trusted-broker system includes only the sensitive identifier attribute values, which can be stored separately (in a dual-database architecture) from FERPA-protected education attribute values. The individual identities would be known only to the trusted-broker system and the agency supplying the identity attribute values. Additional attribute values of interest in multi-agency studies would be provided without identifying the individuals. Two or more agencies would be able to collaborate on research without first aggregating the individual records, provided the identifying attributes have been removed through the trusted-broker system. This procedure allows for more detailed research than other methodologies which would first require aggregation of records.

2.6 Risks and Benefits

The primary risks associated with this research are related to the accuracy and consistency of the identity resolutions in the resulting system. It is understood by all of the authors cited in this research that record linkage and identity resolution are difficult, and the undesired results of false dismissals and false positives are ultimately inevitable. Limiting these identity resolution problems is a key factor in achieving the stated goal of increasing the capability of multi-agency research efforts.

Some technological risks are associated with this research. The volume of data will eventually exceed tens of millions of records spanning several years. Additionally, the system resulting from this research will be intended to continue to function into the future. As research efforts increase to include other state agencies, the volume of records

could escalate quickly, and the limitations of the database system, storage, and processing capabilities may be strained at some point. If so, additional resources may be needed to ensure continued future successes.

Because of the nature of this information and the FERPA requirements, this data should be secured and handled appropriately. A trusted-broker system still requires some transfer of confidential information in the initial phases of identity resolution. Risks are associated with the transfer of this sensitive data, though risks are predictably minimized by the FERPA-compliant capabilities proposed such that data which has been previously resolved is rendered no longer personally-identifiable.

The increase in the capability of agencies to conduct research into data records at the individual level is the key benefit of this proposed research. Currently, the department of education does not utilize all of the proposed methods of identity resolution in this research. As a result a portion of students' longitudinal data is often unavailable to agency researchers. When this portion occurs as unlinked records, it is often removed from consideration in reports. There is a risk that this subset of "problem records" is not a representative sample of the full universe. The reasons for the linkage problems may be related to the students' education deficiencies, either in terms of poor attendance, high mobility, or low quality record-keeping for particular local education agencies. An increase in the overall record linkage for statewide longitudinal education data may have an impact on the research results, perhaps in the form of lowered assessment averages. This risk is not anticipated to be significant but should not be overlooked as a possibility.

CHAPTER 3

ENTITY REFERENCE TABLE

3.1 Initial Population of the Entity Reference Table

The reference data to be utilized in the knowledgebase will be sourced from the official enrollment records of all public schools in the state of Arkansas. Beginning with the most recently completed school year, 2008-2009, the attributes to be utilized for student identity resolution will be Social Security Number (SSN), date of birth (DOB), first name, last name, and a numeric representation of the school district (local education agency – LEA).

These five attributes are de-duplicated across years by design when they are entered into the entity reference table (ERT). In addition to the five attributes sourced from the reference data, the design of the entity reference table also includes attributes for:

- a primary key, randomly assigned identifier value
- record data source(s), indicative of the reference data set(s) providing the five identifier attribute values for the ERT record
- the consolidated identifier key, the lowest primary key value for a particular resolved identity
- the consolidation method(s), indicative of the rule(s) utilized to consolidate multiple ERT records as the same identity

For the first reference file, the default value for the record data source is assigned as “S19” for all records in the newly-created ERT, representing “Student (enrollment data

for fiscal year) 19". Fiscal year 19 corresponds to the 2008-2009 school year. The default value of the consolidated identifier is the randomly assigned primary key value. At this point, no consolidation method has been used and the consolidation method field remains null.

With the introduction of the second reference file to the ERT, the first step is to determine whether the exact combination of the five attribute values for each record have already been included in the ERT by the first reference file. The second reference file is selected as the year prior to the first enrollment file, in this case, the student enrollment file for the 2007-2008 school year. An exact match on all five attributes is performed, and for each match the record data source(s) attribute of the ERT is updated to reflect the additional data source. The value of "S18" is appended to the existing data source value. "S18" is indicative of the student enrollment data for fiscal year 18 (2007-2008).

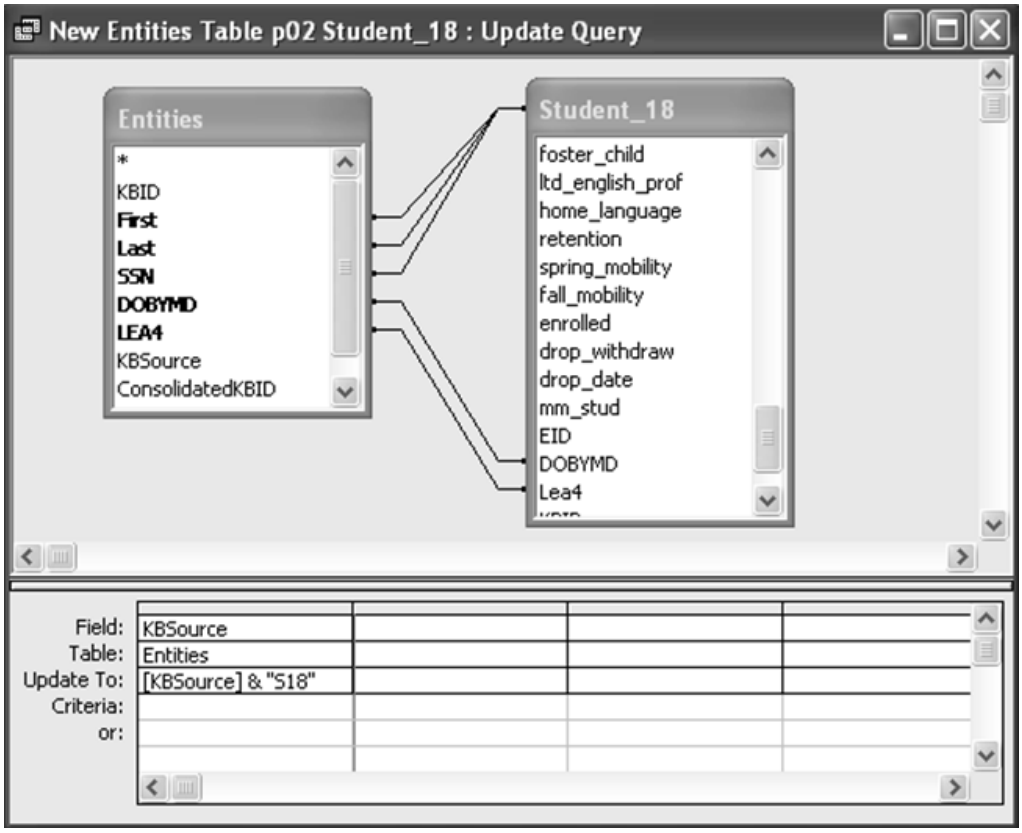

Figure 3. Graphical user interface visualization of ERT source append.

Any records contained in the second reference file which do not already appear in the ERT need to be inserted into the ERT. In order to ensure that there are no duplicate records in the ERT, a non-duplication index should be created for the five fields of SSN, DOB, first name, last name, and LEA. Once this index is created for the ERT, all of the records from the second reference file can be inserted into the ERT because only those records which represent new combinations will be added. Any duplicate records will be disallowed by the index. Alternatively, the second reference file could first be matched to the ERT, updating a "match flag" (which would need to be added to the design of the second reference file) for any records already appearing in the ERT. At this point, only those records of the second reference file which are not flagged would be inserted into

the ERT. In the event that the database platform does not allow indexing without duplicates or in the event that the database platform does not allow appending of records if the index rule is violated, this alternative approach should suffice.

As the new records are inserted into the ERT from those records of the second reference file which not already present, the value for the record data source can be set to "S18" during the insert step. The value for the consolidated identifier will be defaulted to the randomly assigned primary key, but only after the records have been inserted and the primary key has been assigned. This constitutes an extra update step in the ERT build process and is not necessary to perform until just prior to the consolidation process.

It is possible to begin consolidation of the ERT at this point, however, the consolidation process would need to be repeated for each subsequent reference file to be included in the build of the ERT. The addition of each reference file prior to the consolidation of the entities of the ERT is a more practical approach.

Following the same procedure as with the second reference file, the third and subsequent reference files can be added to the ERT. Simplifying the ERT build process to only two steps, the method includes:

1) Determination of those records in current reference file to be incorporated which are already present in the ERT, including an update of the record source attribute.

2) Addition (insertion) of new records from the current reference file which are not already present in the ERT with the appropriate value assigned to the record source attribute.

This two-step process can be repeated until all source data files for the ERT have been incorporated. Once the ERT is populated with all available reference data, the

process for consolidation of entities within the ERT should begin. Detailed SQL statements for initial population of the ERT are included in Appendix A.

3.2 Details for ERT Attributes

For each attribute, it is possible that alternate values could be included in the reference files within the same year or across multiple years. Reference attribute value variations are best understood with the addition of domain knowledge. In longitudinal student enrollment data, the five attributes of particular interest will be the general answers to five key questions about identity:

- Who?
- What?
- Where?
- When?
- How?

More specifically, these questions can be stated as:

- Who are we trying to identify? (A particular student)
- What are the names used by this individual? (First name, last name)
- Where is this individual located? (LEA, local education agency)
- When was this individual born? (Date of birth)
- How is this individual normally identified? (Social Security Number)

Domain-specific details about the five attributes mentioned here (Social Security Number, date of birth, first name, last name, and LEA) and the possible value variations for each are provided in the next five sections of this document.

3.2.1 Social Security Number (SSN)

Social Security Numbers (SSN) are nine-digit numeric values are supplied by the U.S. government's Social Security Administration [SSA] (SSA, 2010). Legitimate values do not begin with 000, 8, or 9. No two individuals should have identical SSN values.

Additional domain knowledge includes:

- (common mistakes) students may have SSN values actually belonging to siblings or parents as a result of errors during enrollment, transpositions. In the event that an SSN value is in use by two or more individuals, at least one error has occurred because only one individual is assigned an SSN value. The errors introduced inadvertently by parents, such as providing the same SSN value for siblings may be only a small portion of a much larger problem with SSN provisions. The discovery of two individuals using the same SSN value is indicative of an error, however, an individual using an incorrect SSN value without a conflict might remain completely undetected and occur more frequently.
- (district problems) one or more districts have used locally-assigned values instead of SSN when student SSN values are unavailable or protected by local rules of anonymity. These locally-assigned values may be unique to a student statewide, or they may only be unique in a particular school district. For example, multiple districts have assigned students sequential values such as 10000001, 100000002, etc, or in reverse, such as 999999999, 999999998, etc. Another district has assigned random SSN

values beginning with 9 which do not appear in any other district records, though it is possible that overlap might occur by chance.

3.2.2 Date of Birth (DOB)

Date of birth (DOB) is a standardized value for the year, month, and day of a student's birth. Though it is a date value, the DOB is stored in a standardized text format. The standardized format utilized in this research is YYYYMMDD (four-digit year, two-digit month, two-digit day), which allows for accurate ascending and descending date sorting of the DOB field not possible with more common display formats (Rud, 2001). Alternate date value formats such as "January 1, 1995" do not sort alphabetically in the correct order of months, since the fourth month "April" occurs first. Numeric formats such as MM/DD/YYYY also sort incorrectly because all DOB values for a particular month are grouped together regardless of the year. For example, "1/1/1995" is sorted closer to a day from the next year, "1/1/1996" than to the next day of "1/2/1995". If desired, the DOB value can be stored as a date data type in the selected database platform, however, it may be necessary to choose a default month or day value in the event that a student's full date of birth is not known. In the YYYYMMDD text format suggested, unknown values may be standardized in storage as zeroes or spaces.

Additional domain knowledge includes:

- (common mistakes) data entry errors at the school districts may transpose the month and day values, international dating formats are already reversed for month and day values, April 5th represented as 5/4 instead of 4/5. Two-digit year entry may result in the wrong century, i.e., data entry

of 95 may result in 2095 instead of 1995 through some erroneous extract, transform, or load of the data.

- (current year) It has also been observed that the year of birth is often incorrectly entered with the current year. A student born on April 5, 1998, may have been entered for the 2007-2008 school year with a date of birth value of 4/5/2007.

3.2.3 First name

First name is the first (given) name of the student as recorded in the student enrollment data sources.

Additional domain knowledge includes:

- (common mistakes) data entry errors at the school districts result in misspellings, unintended additional characters, and inconsistent punctuation such as hyphens or the use of apostrophes to indicate syllable stresses.
- (preferences) students may change their name preference throughout their school years or when changing school districts. First names may be recorded as nicknames or aliases (Billy vs. Bill vs. William) or students may favor their middle names (Neal vs. Sammie).

3.2.4 Last name

Last name is the surname or last (also known as family) name of the student as recorded in the student enrollment data sources.

Additional domain knowledge includes:

- (common mistakes) data entry errors at the school districts result in misspellings, unintended additional characters, and inconsistent punctuation such as hyphens or the use of apostrophes to indicate syllable stresses.
- (district-specific notations) Some school districts have been observed to include additional characters in the last name fields to indicate students who participate in particular programs or have particular characteristics which the administration wishes to note quickly on their computer screens. For example, one district has included an asterisk at the end of the last name field for those students who have special legal instructions regarding guardian limitations for checking the student out of school.
- (legitimate data value changes) Students may experience last name changes due to occurrences such as marriage, divorce, adoption, or other family/guardian events.

3.2.5 Local Education Agency (LEA)

LEA, the local education agency, is notated by a four-digit state-assigned value for each school district. This field is a numeric value stored as text because some values begin with zero and are truncated if stored as a number.

Additional domain knowledge includes:

- (legitimate data value changes) the closing and consolidation of school districts has resulted in the cessation of usage for some LEA values. Historical records may legitimately contain values for closed or consolidated school districts, however, the values have been updated in

more recent records to reflect the current school district LEA for each student impacted.

- (expected data value changes) unlike SSN, DOB, first name, and last name, the LEA value is expected to change for every student who enrolls in a different school district than previous years. The other four attribute values may change, but are not required to change when the LEA value changes.

3.3 Planning Consolidation of the ERT

The goal of the identity resolution and consolidation process is to consistently identify each student even when the attribute values are variable. As the student attribute values change either legitimately or as a result of errors across the longitudinal student enrollment data sources, a consistent identifier should be utilized for each student. Determination of the identity of students is driven by the various record linkage and entity resolution processes. Because of the rules utilized to build the ERT, there are no records which will match on all five attributes.

The first step in ERT consolidation is the deterministic identification of records which have exact matching for four of the five attribute values. This step involves the SSN, DOB, first name, and last name attributes. Matching records represent students who have attended more than one school district (multiple LEA values) in the timeframe of the ERT reference files. To consolidate these records, the SSN, DOB, first name, and last name fields are grouped while the lowest (minimum) value of the randomly assigned primary key identifier is used to represent the group of (consolidated) records. Utilization of the minimum value is somewhat arbitrary since the maximum value could

have also served the same purpose. The important aspect of the methodology is that all consolidation steps will follow the same rule for selection of the group's consolidated identifier. Consistency in this aspect will ensure that logical errors are not introduced into the ERT consolidation results.

After first creating a temporary table containing SSN, DOB, first name, last name, and the minimum primary key, this temporary table is joined to the ERT on the first four attributes and the consolidated identifier is updated for all matching ERT rows with the minimum primary key value, maintaining the convention established in the prior consolidation step. Additionally, the ERT consolidation method attribute should be updated with an indication that impacted records were consolidated utilizing an exact match on four of the five attributes (not including the LEA). One possible way to indicate this consolidation in the method attribute would be to update the method attribute value to "CSDFL", meaning "consolidation by SSN, DOB, first name, and last name". Alternatively the consolidation might be indicated by a value such as "C01", meaning "consolidation one", with the leading zero(es) included to facilitate meaningful metrics and sorting for the consolidation method attribute in the future. A consolidation method translation table is beneficial to describe the consolidation rules in greater detail.

The consolidation steps continue, allowing for a substitution or change to the matching rules which will make it possible for more records to be consolidated. Each rule substitution or change is intended to "loosen" the attribute value requirements slightly when compared to the prior rule, but to maintain the confidence that the resulting consolidations accurately apply to the same individual identities. As mentioned in the attribute details from the prior section, common mistakes and value changes may be

expected to occur in multiple records for the same individual. These value differences will often prohibit exact matching of attribute values, and will need to be allowed via translation tables or approximate matching techniques in order to facilitate additional consolidations. Proposed consolidation rules are outlined in the table below. These rules do not represent an exhaustive set of possibilities. Given the nature of the match types and the five attributes, an unwieldy number of rule combinations are possible, however, domain expertise allows the researcher to more quickly focus on the particular combinations which will be of value to the methodology being proposed (Summers, 2006). Selected SQL statements for consolidation of the ERT are detailed in Appendix B. Note: The qTR notation referenced in the table will be discussed in the next section of this research.

Table 5 Proposed consolidation rules itemized by attribute matching characteristics

Rule	DOB	FN	LN	SSN	LEA	Comment
1	X	X	X	X		Change of school district (LEA)
2	X	X	X		X	Change of SSN within LEA
3		X	X	X	X	Change of DOB within LEA
4	X		X	X	X	Change of First name within LEA
5	X	X		X	X	Change of Last name within LEA
6	X	LN	FN	X		Reversed First and Last names
7	X	q	X	X		Exact D, L, S, and qTR First names
8	X	X	q	X		Exact D, F, S, and qTR Last names
9		X	X	X		Exact F, L, S
10	X		X	X		Exact D, L, S
11	X	X		X		Exact D, F, S
12	X	X	X			Exact D, F, L
13	X	q	q	X		Exact D, S, and qTR F, qTR L
14	X	q		X		Exact D, S, and qTR F
15	X		q	X		Exact D, S, and qTR L
16	q	U	U	q		Unique F+L combo, qTR D, qTR S
17	q	q	q	X		Exact SSN, and 2 out of 3 qTR D,F,L

Translation tables may be utilized or created to facilitate additional matching for differing values. The first example of this type of table would be a nicknames table, containing commonly known variations of individual first names. Researching common nicknames for first names in the United States reveals that there are no guidelines for which names may be considered nicknames. Over time first names may fluctuate in popularity (frequency of use), while some nicknames are common to multiple names, such as "Chris", which may be a nickname for "Christopher", "Christian", "Christina", or may not be a nickname at all if the individual's official first name is "Chris". Care should be taken not to conclude that "Christopher" is a valid nickname for "Christina" simply because both can be referenced by the nickname use of "Chris". It may be desirable to incorporate a table of nicknames into the consolidation process, however, the results of doing so will vary according to the breadth and depth of the table used.

Another obstacle for incorporating a translation table may be the lack of any existing research into the possible variations in attribute values. While the most common nicknames for first names may be well documented, less frequently used first names may not be included in existing nickname tables. Other attribute value variations may be due to domain-specific (or field-specific) reasoning such as the closing of school districts, which are subsumed by other districts. References to the prior school district should be considered equivalent to references to the new district for the same individual at an earlier period in time. Variations which occur for cognitive, phonetic, or typographical reasons may not represent actual nicknames or aliases for data values if they are unintended errors. It is unlikely that any table of alternative values could predict all possibilities. To overcome the limitations of translation tables, approximate string

matching (pattern matching) may be implemented to increase the potential consolidation candidates.

3.4 Approximate String Matching

3.4.1 Background

String matching is based upon exact matching techniques, most often associated with database statements written in SQL, either in the form of the GROUP BY clause or the WHERE String1=String2 expression. This type of matching is valid but incomplete. The text string "NICK" cannot be matched to "NICHOLAS", though the two names are often interchangeable for the same person. A table of known aliases or nicknames may be available which could allow for NICK-to-NICHOLAS matching, however, these tables are always incomplete and do not normally include transpositions, such as "JOHN" to "JONH". Infrequently used names are not likely to be present in a standard nickname table, though there may be multiple variations of the name for the same individual.

Furthermore, nickname or alias tables are domain-specific and often field-specific, which is an undesirable requirement if attempting any general-purpose text comparisons. When the strings being compared are alphanumeric, the concept of "nicknames" may have no meaning. For example, consider "T8R9X" compared to "TBR9X". It may be possible to create a table of aliases for domain-specific alphanumeric values, however, it is unlikely that all possible errors could be anticipated in advance.

Standard approaches to approximate string matching include both Edit distance and Soundex. Soundex immediately fails when two strings do not begin with the same

character (National Archives, 2007). Additionally, Soundex was designed for words that "sound alike", and not words which may be typed incorrectly. The transposition of two consonants in a string are likely to result in two different Soundex values, which makes approximate matching very difficult.

Edit distance, in its most basic form, returns an integer value representing the minimal number character insertions, deletions, or substitutions necessary to transform one string into the other (Gilleland, 2009). By its definition, edit distance is not sensitive to the position of where string differences occur. For example, all pairs of strings differing in only one character all return an edit distance of 1 (Hall & Dowling, 1980). Weightings may be incorporated into the edit distance formula to differentiate results by position, but this process increases complexity even further (Damerau, 1964).

3.4.2 q-Gram Definition

The term 'q-gram', also called 'n-gram', refers to a subsequence of q items from a given sequence (Christen, 2006). With respect to data values such as the earlier examples of the text value "CHRIS", all possible q-grams where q = 3 would be the subsequences "CHR", "HRI", and "RIS". The value of q can be any number between 1 and the length of the sequence. A variety of existing methods for pattern matching utilize q-grams in their approach (Gravano, 2001).

3.4.3 Tetrahedral Numbers

A tetrahedral number, also called a triangular pyramidal number, is a figurate number corresponding to the number of discrete points arranged into a tetrahedron (triangular base pyramid). Calculation of a tetrahedral number follows the formula:

Equation 1 Tetrahedral number (Weisstein, 2010)

$$T_n = \frac{n(n+1)(n+2)}{6}$$

For example, the tetrahedral number calculated for n = 4 is Tn = 20, illustrated as 20 discrete points in Figure 4.

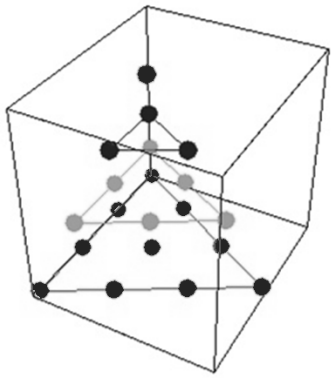

Figure 4. A tetrahedral arrangement for Tn with side length n = 4, represented by 20 discrete points.

Consider the text string “JOHN”. Possible q-grams for “JOHN” are “J”, “O”, “H”, and “N”, when q = 1; “JO”, “OH”, and “HN”, when q = 2; “JOH” and “OHN”, when q = 3; and “JOHN” when q = 4. These 10 subsequences constitute all possible q-grams for “JOHN”. If each letter in these ten subsequences is considered a discrete point, there are 20 discrete points for the text string of length 4, identical to the calculation of Tn for n = 4.

A triangular arrangement can be used to illustrate all possible subsequences for a pattern (string of characters) of any length. In each case, the length n of the pattern will represent the number of subsequences on each side of the triangle, the number of subsequences in total will be Σn, and the total number of characters will be Tn. Figure 3 has a triangular shape with each side consisting of 4 subsequences, there are 10 total subsequences, and the number of discrete points (individual characters) is 20. Though the display is two-dimensional, the 20 discrete points could theoretically be positioned to match Figure 4.

Figure 5. Triangular arrangement of subsequences for "JOHN".

Combining the q-gram subsequences for character strings with the mathematical aspects of tetrahedral numbers allows for a comparison of any two string patterns. Rather than selecting a particular value for q, this more comprehensive approach would include all possible values for q.

3.4.4 Combined Approach to Comparisons

To compare any two strings, S1 and S2, the comprehensive q-gram subsequences for S1 are located within S2, if possible. Suppose that S1 = “JOHN” and S2 = “JONH”. Each of the ten possible q-gram subsequences in Figure 5 for “JOHN” are indicative of pattern similarity if those subsequences can also be located as a subsequence of “JONH”. Figure 6 highlights the subsequences from Figure 5 in common with subsequences of string “JONH”.

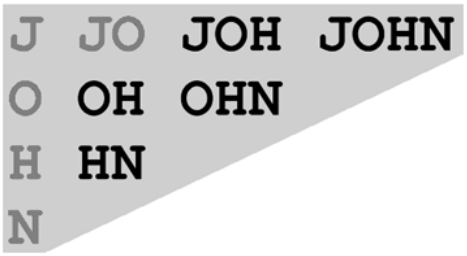

Figure 6. Highlighted subsequences of "JOHN" which are shared with subsequences of "JONH". {"J","O","H","N","JO"}.

In order to determine the similarity of two strings (S1 and S2) utilizing the tetrahedral aspects of q-gram subsequences, Tn is calculated where n is the length of S1, and Q is the count of discrete points (characters) shared by the q-gram subsequences in common to both S1 and S2.

The simple ratio qTR can be calculated as:

Equation 2 Simple q-gram tetrahedral ratio (qTR)

$$qTR = \frac{Q}{T_n}$$

Given that $0 \leq Q \leq Tn$, it is true that $0 \leq qTR \leq 1$. For the example where S1="JOHN" and S2="JONH", qTR = 0.3. This qTR value is obtained because:

Q = 6, as illustrated by the 6 discrete points (individual characters) in the 5 subsequences shared by S1 and S2, highlighted in Figure 6

n = 4, the length of S1= "JOHN"

Tn = 20, the tetrahedral number when n = 4, and the total number of discrete points (individual characters) in Figures 5 and 6

qTR = Q / Tn = 6 / 20 = 0.3

Another example mentioned in the introduction of this research was that of “T8R9X” and “TBR9X”.

In this example, qTR = 11 / 35 = 0.314. There are 11 characters in the 7 subsequences highlighted in Figure 7 reflecting the same process used for Figure 6, allowing S1=”T8R9X” and S2=”TBR9X”.

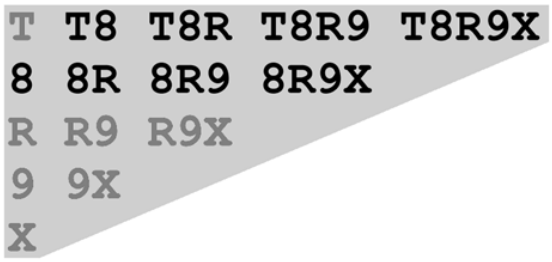

Figure 7. Highlighted subsequences of "T8R9X" which are shared with subsequences of "TBR9X". {"T","R","9","X","R9","9X", "R9X"}.

3.4.5 Coded Implementation

Despite the comprehensive q-gram aspects of the qTR approach, the coded implementation of qTR is surprisingly simple. Utilizing Visual Basic for Applications (VBA), the implementation of the proposed qTR function would consist of only a dozen lines of code as a simple nested loop.

```
i = 1
Do While i <= Len(S1)
    j = 1
    Do While j <= Len(S1) - i + 1
    Tn = Tn + j
        If InStr(S2, Mid(S1, i, j)) > 0 Then
        Q = Q + j
        End If
    j = j + 1
    Loop
i = i + 1
Loop
qTR = Q / Tn
```

Figure 8. Visual Basic for Applications (VBA) implementation of simple qTR.

The qTR implementation would also include the function declaration including the input variables S1 (String 1) and S2 (String 2) as well as the declaration of integer variables i, j, Q, and Tn with initial values of zero. Though other modifications or error checking may be optionally included, no other code would be necessary to implement qTR.

It is possible to obtain a qTR result which is applicable even if the order of S1 and S2 values are reversed. “JOHN” and “JONH” can be interchanged with no effect on the qTR result of 0.3. Likewise, “T8R9X” and “TBR9X” can be interchanged with no effect on the qTR result of 0.314. This string commutability is only a special case of the qTR algorithm, limited to cases where the lengths of S1 and S2 are identical.

In the event that S1 and S2 have differing string lengths n, the respective values of Tn also differ. As a consequence, the qTR result are expected to depend upon the ordering of the two comparison strings. Observe the difference in qTR results when evaluating S1=“NICK” and S2=“NICHOLAS”:

Q = 10 due to the shared subsequences of “N”, “I”, “C”, “NI”, “IC”, and “NIC”

n = 4, corresponding to the length of S1

Tn = 20 when n = 4

qTR = Q / Tn = 10 / 20 = 0.5

Compare to S1=”NICHOLAS” and S2=”NICK”:

Q = 10 due to the shared subsequences of “N”, “I”, “C”, “NI”, “IC”, and “NIC”

n = 8, corresponding to the length of S1

Tn = 120 when n = 4

qTR = Q / Tn = 10 / 120 = 0.083

Because the interchange of the two strings can cause the value for qTR to differ, it may be desirable to devise a method to ensure consistent results for calculating qTR regardless of string ordering.

3.4.6 Adjusted Calculation

The qTR calculation is determined by the values of Q and Tn. The value of Q for any two strings is a constant, determined by the particular subsequences shared regardless of string order. Consequently, the differing values in qTR occur when the length values differ. A simple adjustment for qTR would incorporate both possible values of n in order to eliminate the impact of string order. Utilizing a length-weighted average for the two qTR results effectively produces the desired order-independent effect. If n1 represents the length of S1 and n2 represents the length of S2, an n-weighted average for qTR would be calculated as:

Equation 3 Adjusted q-gram tetrahedral ratio (qTR)

$$qTR(adjusted) = \frac{\frac{n1 \times Q}{T_{n1}} + \frac{n2 \times Q}{T_{n2}}}{n1 + n2}$$

Although the formula appears significantly more complicated than the earlier version, the change to the code is minor, requiring only the introductions and assignments of a few new variables.

```
i = 1
Do While i <= Len(S1)
    j = 1
    Do While j <= Len(S1) - i + 1
        If InStr(S2, Mid(S1, i, j)) > 0 Then
        Q = Q + j
        End If
    j = j + 1
    Loop
i = i + 1
Loop
n1 = Len(S1)
n2 = Len(S2)
Tn1 = (n1)(n1+1)(n1+2)/6
Tn2 = (n2)(n2+1)(n2+2)/6
qTR = (n1*Q / Tn1 + n2*Q/Tn2)/(n1+n2)
```

Figure 9. Visual Basic for Applications (VBA) implementation of adjusted qTR.

Revisiting earlier examples, the adjusted qTR values would be:

qTR = 0.3 for (“JOHN”,”JONH”)

qTR = 0.314 for (“T8R9X”,”TBR9X”)

qTR = 0.222 for (“NICK”,”NICHOLAS”)

The standard approaches to approximate string matching do not demonstrate this precision in result differentiation. For the three examples above, Soundex results in a “match” for (“JOHN”,”JONH”) and a “no match” for the other two cases. Edit distances for the three examples are 1, 1, and 5, respectively.

Summarizing the findings of this research, the proposed qTR methodology:

- utilizes all possible q-gram subsequences for two strings
- incorporates the mathematical concept of tetrahedral numbers
- determines a similarity ratio for any two strings
- is not dependent upon the order of the two strings
- requires minimal code to implement

Additionally, the qTR methodology as described appears to:

- have no limitations to any particular set of characters
- be applicable for both left-to-right (LTR) and right-to-left (RTL) directional text situations, provided both strings are written in the same manner.

3.4.7 qTR Utilization and Recommendations

A minimum qTR value should be determined through subject matter expertise for the intended implementation. In the testing of the qTR associated with this research, a minimum qTR value of 0.25 appears significant as a threshold for approximate matching. The purpose of approximate pattern matching is to increase automated record linkage. Valid linkages will be determined by the user and should represent those "near matches" that the user would approve if doing the comparison work manually. It may be necessary to determine multiple qTR value ranges corresponding to those string comparisons which are deemed highest-confidence, acceptable, unacceptable, or worthy of visual inspection.

Implementation of the qTR for non-Western alphabet/keyboard data would be beneficial to further research the "universal" aspects of the methodology. Subject matter expertise in those languages or data sets would enhance the research immensely. Though

the qTR as described is neither domain-specific nor field-specific, it is understood that the implementation of the qTR to specific applications may be enhanced by domain-specific or field-specific coding adjustments, such as an “extra credit” factor for strings which begin with the same letter. The further enhancements to the qTR may improve the performance of the approximate string matching by incorporating elements from alternate methods, such as the phonetic aspects of Soundex. Comparison and modification related to additional techniques, such as the Jaro-Winkler string comparison (LingPipe 2009). may enhance the utility of qTR. The qTR may be utilized to create nickname or alias tables for a particular implementation if frequently-occurring string combinations are determined to be acceptable as matches.

The use of qTR in non-name approximate matching has not been researched at this point. It is possible that qTR may provide some value to research of other types of character strings or possibly in the comparison of data which is only represented by character strings in electronic forms but in fact refers to real-world objects or images. Additionally, it is likely that the incorporation of existing rules for edit distance or phonetic approximations would produce better results. The complexity of the implementation may increase by these adjustments, however, that complexity may be warranted if the results are also beneficial.

Teacher 1	Teacher 2	Teacher 3
AMON	DELOR	WENDLYN
EMON	DELORIS	WENDOLY
LAMON	DOLORES	WENDY
LEMAN	DOLORES A	WENDYLI
LEMON F		WENDYLIN C

Figure 10. Examples of consolidated first names identified utilizing qTR.

The first name examples displayed belong to individuals whose identifying attribute values provided evidence for identity consolidation despite the differences in first name values. These names are visually similar and can be readily understood to belong to the same individuals, however it is unlikely that nickname tables would contain these particular name values. The utilization of the qTR is one method to identity these types of name variations in knowledge-driven identity resolution efforts.

3.5 Extended ERT Consolidations

Continuing the consolidation process for the ERT, the use of approximate string matching allows a number of additional identities to be consolidated. Though it is possible to continue developing rules for consolidations, it is important to avoid over-consolidations which would identify two different individuals with the same identity. The possibility of under-consolidating is preferable to over-consolidation because it is always possible to consolidate two identities in the future as additional information becomes available, however, it is very difficult to deconsolidate two individuals who have been incorrectly consolidated at a prior date. References to two individuals by the same identifier, creates a problem in knowing which individual was intended by each reference once the problem is identified.

As each step in the consolidation process occurs, it is possible that the consolidated identifier will be inconsistent for portions of the identity group. This occurs when a consolidation rule applies to some, but not all, of the records in a consolidated group.

Consider the example in Table 6.

Table 6 Example records in an entity reference table (ERT), not yet consolidated

Entity ID	Attribute A	Attribute B	Attribute C	Consolidated ID	Consolidation Comment
381	X	23	G	381	None
437	X	24	G	437	None
982	Y	24	J	982	None

As these records are added to the ERT, they are by default given a consolidated identifier which matches their randomly assigned entity identifier. No consolidations have been identified at this point in the process. Consider the first consolidation rule in this example.

Example Consolidation Rule 1 – if Attribute B matches, consolidate the entities.

Following the application of Example Consolidation Rule 1, the ERT would reflect the state shown in Table 7.

Table 7 Example records in an entity reference table (ERT) following a single consolidation step

Entity ID	Attribute A	Attribute B	Attribute C	Consolidated ID	Consolidation Comment
381	X	23	G	381	None
437	X	24	G	437	Rule 1
982	Y	24	J	437	Rule 1

Observe how the consolidated identifiers for the second and third records now match, with a value of 437. At this point we can apply the second consolidation rule.

Example Consolidation Rule 2 - if Attributes A and C match, consolidate the entities.

Following the application of Example Consolidation Rule 2, the ERT would reflect the state shown in Table 8.

Table 8 Example records in an entity reference table (ERT) following two consolidation steps

Entity ID	Attribute A	Attribute B	Attribute C	Consolidated ID	Consolidation Comment
381	X	23	G	381	Rule 2
437	X	24	G	381	Rule 1, Rule 2
982	Y	24	J	437	Rule 1

Despite the correct application of two consolidation rules, which should have theoretically consolidated these three records into the same identity, the resulting ERT shows that the third record does not match the consolidation identifier of the other two records. This is not a problem with the consolidation rules, instead it is an artifact of the situation in which a consolidation rule may impact only part of the previously-consolidated identity group records. A correction step is required to solve this problem.

A self-referential query identifying the lowest consolidation identifier for all records in a consolidation group will successfully update the records which are out of sync. This step is required prior to any attempts of identity resolution for transaction data.

Following the application of lowest consolidation identifier logic, the ERT would reflect the state shown in Table 9.

Table 9 Example records in an entity reference table (ERT) following a third step to isolate unsynchronized records

Entity ID	Attribute A	Attribute B	Attribute C	Consolidated ID	Consolidation Comment
381	X	23	G	381	Rule 2
437	X	24	G	381	Rule 1, Rule 2
982	Y	24	J	381	Rule 1, lowest

Though these examples are simplistic and do not represent the exact situations of the longitudinal student data, these examples demonstrate the problem outcome of unsynchronized consolidation identifiers.

3.6 Discussion of ERT Considerations

3.6.1 Set-specific Considerations

In closed set resolution the goal is one student, one year, one use – when comparing two sets, a single student should appear only once for the year, and should be used only once in a matching pair.

The leftover subset of students is the full set of possible matches – once rules have identified students under consideration, the remaining student subsets are the only possible matches for the remaining unmatched records.

3.6.2 Tuple-specific Considerations

A student should have a contiguous history and single location – except for ETL timing-related issues (the delay that may occur between a school enrolling a student and the prior school dropping the student), a student should not "attend" two schools at once, and should not have gaps in attendance.

3.6.3 Element-specific Considerations

SSN may take multiple forms – depending upon the year, school, and policy, the SSN may be a sequential value, a school-assigned identifier, an actual SSN, or some other value. The actual values for SSN should be recognizable, and alternate values should be understood as non-SSNs.

First names can take multiple forms, for example, Robert is more formal, while Bob and Bobby are common nicknames for Robert. Additionally, students may prefer to use their middle name if they are given the choice, while the school's official records may utilize their first name.

Other domain-specific considerations (in these or other categories) should be identified and documented in order to leverage the subject-matter expertise available for the identity resolution system.

3.6.4 Directional Nomenclature

Initial consideration of the method for resolving identities included a discussion of the differences (pros and cons) in terms of a directional aspect of consolidation.

- Horizontal – two data sets selected from the same year
- Vertical – two data sets selected from different years

Further research into this aspect of consolidations revealed that a more important factor for consolidations involves the difference between data sets which are determined to be authoritarian and to be included in the ERT, as opposed to data sets which are not authoritarian and will not be included in the ERT. The judgment of an "authoritarian data set" is given to an organizational subject matter expert, in the case of education data, the authoritarian data sets include those sets which compile "data of record" with regard to the identity attributes of the students. In the case of the Arkansas Department of Education, these data sets will include the official enrollment records for school years.

Non-authoritarian data sets include those sets in which the data is provided by less-than-reputable sources, in this case, the students themselves. It is understood that data sets which are compiled as a result of students' "bubbling in" of test identification information are quite prone to a number of errors. Many students (though not all) do not take the data collection effort seriously and have been known fill in only partial information or even fictitious identity information. These data sets do represent usage of these identity attribute values, however, they are not authoritarian enough to include in the ERT. Some organizations may find all records of value to the ERT, regardless of the data source, however, the subject matter experts at the Arkansas Department of Education do not hold this view.

With regard to the definitions of horizontal and vertical nomenclature, it is recommended that all authoritarian data sets are included in the ERT, regardless of direction. As it impacts this research, only the vertical direction will be used as a result of one student enrollment data set per school year being included in the ERT. It would be valid to include data of a horizontal nature if more than one data set per year was

determined to be authoritarian (Calvanese, De Giacomo, Lenzerini, & Rosati, 2004, and De Giacomo & Lembo, 2005).

3.7 Transaction Data Sets

Transaction data sets in this research consist primarily of the results of standardized assessments of student groups. For instance, the results of the ACT (College Board) are provided by the testing company to the Arkansas Department of Education with identifying attributes provided by each student. Similarly, standardized tests for students in elementary and intermediate schools may demonstrate attribute values provided manually by teachers, testing coordinators, or students. These attribute values are sometimes optional and do not always match the five key attributes of this study.

It is necessary to correctly identify the student entities associated with each transaction event in order to obtain a complete annual and longitudinal record of student data. The goal of this process would be 100% identifications which are both accurate (identifying the correct student) and complete (no identities left unresolved). It is unreasonable to require 100% resolution metrics, but it not unreasonable to keep this ideal goal in mind.

The reason that it is unreasonable to require 100% resolution metrics is that the input data sets may contain records which have no identifying information. It is not unusual for some test result records to contain scores, but no names or other identifiers. This may occur if the identification portion of the test is accidentally left blank. Generally speaking, the testing environment continues to utilize “bubble sheets” which are filled in by hand with a No. 2 pencil. The testing company provides instructions for

the proper methods of completing the sections however there is no guarantee that each student will attentively follow those instructions.

Additionally, transaction data sets from the ACT often include records belonging to prospective college students who are not currently in the Arkansas Department of Education reference data. For example, a prospective college student aged 27 who takes the ACT may correctly identify his former high school, however, he may not have attended that high school in many years. It is possible that the date of birth attribute can assist in the identification of records which will be unresolved due to the age of the student however the date of birth attribute is not always populated.

3.8 Rules for Identity Resolutions

As in the consolidation process, the steps taken to increase the number of consolidations are echoed in the identity resolution process. Beginning with the most complete matching possible for each identity attribute, matches to the ERT result in a positive identification of the student. Following the same procedures as consolidations, attribute matching is loosened or adjusted to allow for less stringent matching which is still considered high-confidence.

Unlike the consolidation process, identity resolution continues when matching rules fall below the confidence requirements of consolidation. For example in the event that an identity resolution is requested based solely on a Social Security Number (SSN), assuming all other attribute values are null, a positive match to the ERT on SSN alone can result in a resolution. It is still important to determine whether the SSN is a valid value (according to the rules of the Social Security Administration), and also to determine whether multiple entities of the ERT have been associated with the SSN. In the event

that either the SSN is of an invalid form or the ERT has multiple identities associated, the resolution should indicate this lower level of confidence in order for the user to be able to make an informed decision about record usage. Regardless of SSN or ERT results, identity resolutions should include both an identifier and a resolution type (or descriptive) for each record.

Resolution rules are adjustable from a bottom-up approach to creating this type of methodology. The advantage of the bottom-up approach is not only speed of development, but the ability to modify and supplement the identity resolution functionality as knowledge increases (Dyché & Levy, 2006). The identity resolution rules proposed in this research constitute a combination of the same rules utilized in consolidation and new rules which are not suitable for identity consolidation but which do have value in the event that other rules have failed and transaction records remain unresolved. These rules are not intended to be exhaustive of all possibilities, but represent a subset of rules capable of resolving identities for transaction data sets.

Table 10 Proposed identity resolution rules itemized by attribute matching characteristics

Rule	DOB	FN	LN	SSN	LEA	Comment
1	X	X	X	X	X	Exact match on all 5 attributes
2	X	X	X	X		Change of LEA
3	X		X	X		Exact D, L, S
4	X	X		X		Exact D, F, S
5	X	X	X			Exact D, F, L
6	X	LN	FN	X		Reversed First and Last names
7		X	X	X		Exact F, L, S
8	X			X		Exact D, S
9			X	X		Exact L, S
10			q	X		Exact S. qTR L
11	X	X	q			Exact D, F, qTR L
12	X	q	X			Exact D, L, qTR F
13	q	X	X			Exact F, L, qTR D
14	md	X	X			Exact F, L, and match month/day
15	q		X	q		Exact L, qTR D, qTR S
16	y	U	U		X	Exact LEA, unique D+L combo
17	U		U		U	Unique D+L+LEA combo
18	y		U		U	Unique L+LEA combo, match year
19		q	X	i	X	Exact L, LEA, qTR F, invalid S
20				U	U	Unique S+LEA combo

3.9 Longitudinal Aspects

The longitudinal aspects of the education data create a logical timeline of student identity which can be utilized by the identity resolution methodology to ensure historical completeness. As described in an earlier section, the fact that the entities in this research represent actual students requires that the student "exists" in exactly one place at all times (with only slight gaps or overlaps due to various source database updating procedures). Using the ERT to define the longitudinal aspects of each resolved identity can isolate those entities who have "missing time" in their education histories. These temporal gaps can be investigated to determine causes which may indicate omissions in reporting,

absenteeism, source data errors, and potential improvements in the consolidation methodology.

Whenever an entity's resolved records display a longitudinal gap determined to be attributable to a possible cause, the state agency may have the ability to contact the representatives of the local agency to resolve the issues. Because students in this longitudinal data system are expected to be located continuously for grades 1 to 12 (if not also including kindergarten), gaps which are identified at the state level prior to graduation should represent a real-world opportunity to better understand the situation at the local level. The results of these investigations should indicate the appropriate course(s) of action to improve record-keeping and continuous longitudinal resolutions at the local level for state reporting in the future.

As described in Chapter 1, the purpose of the rules planned for entity reference table consolidations and identity resolutions are to emulate the decision-making process of a knowledgeable person who is tasked with determining whether the identities of two records in fact refer to the same real-world entity. These rules should be adjusted whenever it is determined that the results do not match those intended. The domain expertise of researchers in the agency utilizing these methodologies should be understood in order to more effectively mimic the identity-resolving capabilities of those individuals who would otherwise manually apply the best practices for the organization. The automation of these best practices serves to improve the speed of the application, but does not imply any allowed reduction in the quality of the work.

CHAPTER 4

RESULTS OF ACTUAL PROCESSING

4.1 Initial Load of the ERT

The number of records in the ERT following the initial load of the student enrollment tables for four years (2005 to 2009) is 879,780. Exact matching on all five attributes eliminated duplication of 1,346,808 records. Of the 879,780 records in the ERT, 259,825 appeared in all four years of student enrollment data (29.5%). A total of 255,778 records appeared in only one of the four sources (29.1%). The remaining 41.4% of records in the ERT appear in two or three years of source data. Additionally, of the 624,002 records appearing in two, three, or four years of data, only 9,633 records display non-consecutive year timeframes (just over 1%). This small number approximates the number of students who returned to a school district after attending school elsewhere at some point within the four years of data. SQL statements utilized in the load of the ERT records from the four years of student enrollment data are included in Appendix A.

Table 11 Description of four years of student enrollment data sets

Data Set	Quantity
Student Enrollment 2005-2006	590,806
Student Enrollment 2006-2007	584,098
Student Enrollment 2007-2008	588,279
Student Enrollment* 2008-2009	463,405
TOTAL	2,226,588

* database changes in 2009 impacted quantity

These 2,226,588 records were loaded into the ERT utilizing the rule that any exact match on all five attributes resulted in only an update to the record source notation,

rather than the addition of a new record in the table. As a result, 879,780 records were inserted into the ERT, with data source notations as shown below.

Table 12 Resulting data source attribute values following the initial population of the ERT

ERT Records Data Source(s)	Net Records	Percent	Comment
S19S18S17S16	259,825	29.5%	4 years
S16	109,805	12.5%	1 year
S18S17S16	106,998	12.2%	3 years
S17S16	88,069	10.0%	2 years
S19	76,714	8.7%	1 year
S19S18	62,099	7.1%	2 years
S19S18S17	57,105	6.5%	3 years
S18	48,904	5.6%	1 year
S18S17	40,273	4.6%	2 years
S17	20,355	2.3%	1 year
S19S17S16	3,031	0.3%	3 years*
S19S18S16	2,092	0.2%	3 years*
S18S16	2,015	0.2%	2 years*
S19S16	1,499	0.2%	2 years*
S19S17	996	0.1%	2 years*
NET TOTAL	879,780	100.0%	

* data sources display at least one year of gap

Though students regularly change school districts, the number of students who subsequently return to a prior school district is very low. These cases are shown in the table as data sources which display at least one year of gap prior to a “return” to the same school district (LEA) in a future year, totaling 9,633 records (just over 1%).

4.2 Consolidation of the ERT

4.2.1 Deterministic Rules

Implementation of the proposed consolidation rules (Table 5) provides a real-world opportunity to evaluate the utility of each step. Each of the seventeen rules is performed on the longitudinal education data available for this research and described individually in the succeeding sections.

The first rule of consolidation requires exact matching for four of the five attributes (date of birth, first name, last name, and SSN). The local education agency (LEA) is not required to match. These records are indicative of students who changed school districts at some point in the four year span of the data sources. Prior to the implementation of this step, the number of student identities in the ERT was 879,780. Following this consolidation step, the number of student identities became 741,044, indicating that a change of school district (only) occurred for 138,736 records, just over 15% of the ERT.

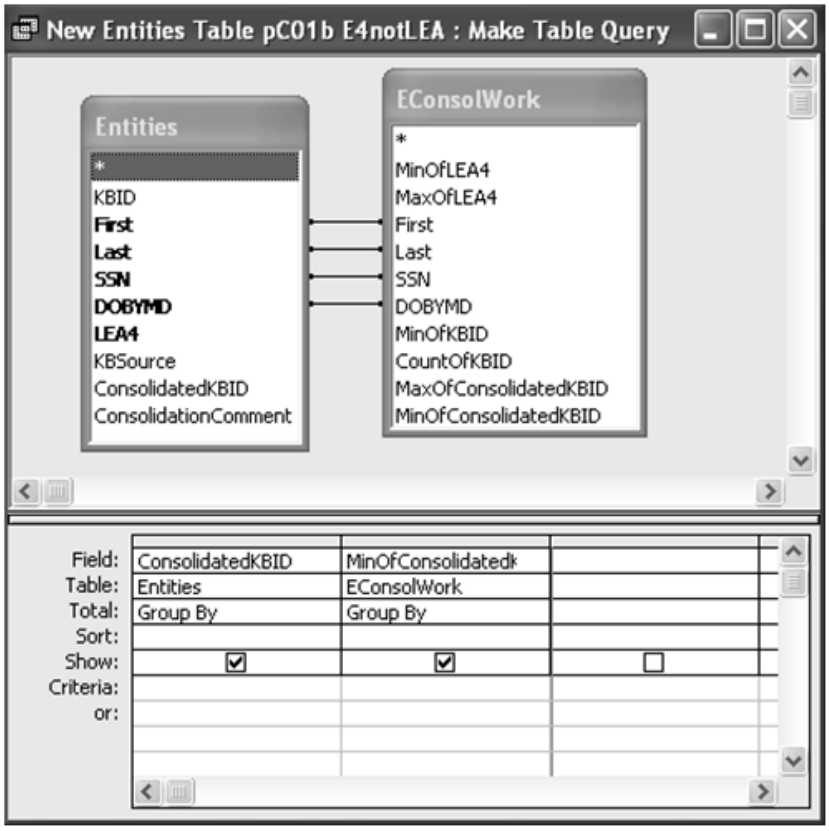

Figure 11. Graphical user interface visualization of the second rule of ERT consolidation.

The second rule of consolidation requires exact matching for four of the five attributes (date of birth, first name, last name, and LEA). These records are indicative of some type of change in the Social Security Number as recorded by a single school district. These changes may occur when a temporary SSN is assigned until the actual value can be determined, or in cases where the school district has a policy to use some value other than actual SSN as the SSN value. Following the implementation of the second consolidation step, the number of student identities changed from 741,044 to 730,190, indicating that a change of SSN value (only) occurred for 10,854 records (about 1.2% of the ERT) for two or more records within the same school district. Of the 274

school districts in the ERT, 159 had SSN value changes for 2% or more of their total ERT records. Eight school districts had SSN value changes for 5% or more of their total ERT records. One school district had 27.3% SSN value changes. This particular school district had replaced SSN with randomly assigned numbers beginning with '9' during one of the four years of data.

The third rule of consolidation requires exact matching for four of the five attributes (first name, last name, LEA, and SSN). These records are indicative of some type of change in the date of birth as recorded by a single school district. These changes may occur due to typographical errors or defaulted date values, which are later corrected in the enrollment data. Following the implementation of the third consolidation step, the number of student identities changed from 730,190 to 725,990, indicating that a change of the date of birth value (only) occurred for 4,200 records (about 0.5% of the ERT) for two or more records within the same school district.

The fourth rule of consolidation requires exact matching for four of the five attributes (date of birth, last name, LEA, and SSN). These records are indicative of some type of change in the first name as recorded by a single school district. These changes may occur due to typographical errors which are later corrected in the enrollment data. Changes may also occur when a student chooses to be known by a middle name, or if two schools in the same district follow different rules regarding the use of middle initials in the student first name attribute value. Following the implementation of the third consolidation step, the number of student identities changed from 725,990 to 719,811, indicating that a change of the first name value (only) occurred for 6,179 records (about 0.7% of the ERT) for two or more records within the same school district.

The fifth rule of consolidation requires exact matching for four of the five attributes (date of birth, first name, LEA, and SSN). These records are indicative of some type of change in the last name as recorded by a single school district. These changes may occur due to typographical errors which are later corrected in the enrollment data. Changes may also occur when a student's last name changes due to marriage, divorce, adoption, or if two schools in the same district follow different rules regarding the use of hyphenated names in the student last name attribute value. Following the implementation of the third consolidation step, the number of student identities changed from 719,811 to 713,672, indicating that a change of the first name value (only) occurred for 6,139 records (about 0.7% of the ERT) for two or more records within the same school district.

Upon completing the first five rules of consolidation, all possible consolidations including exact matching for four of the five attributes have been identified. The net effect is that the ERT table containing 879,780 records is reduced to 713,672 consolidated identities, a reduction of 18.9%. Though the order of consolidation has been described as first rule, second rule, etc., the non-overlapping nature of the rules to this point are order-independent. These five rules could have been implemented in any order and the results would be the same. Identity consolidation continues utilizing variations on exact matching and inexact (fuzzy, approximate) matching for the five attribute values.

4.2.2 Semantic Reconciliation Rules

Given the matching type variations for each attribute of omission, exact matching, transposition, qTR, nicknames, and other potential near-match algorithms, an exponential number of possible consolidation rules can be described for these five

identity attributes. While only a small percentage of these possible rules would result in a confident match, it is difficult to isolate the particular rules which are most appropriate to continue this consolidation methodology. Clearly, rules which utilize only one attribute are ineffective for identity resolution. For example, a rule which would suggest that two identities are the same if they share a first name only, or if they share a date of birth only, would result in an extreme number of over-consolidations.

Because all possible rules utilizing exact matching with four of the five attributes have been implemented, it is logical that the next rules should utilize exact matching with three of the five attributes and potentially include approximate matching techniques on the other two attributes. It is at this point in the process that subject-matter expertise guides the researcher to determine appropriate next steps. As mentioned earlier, the key to effective identity resolution is emulating an intelligent user's ability to determine a match based on a variety of factors (Informatica, 2008). While this process is subjective, it is difficult to justify the relegation of consolidation rules to the more objective approaches which would systematically attempt all possible rule combinations. Though these rule combinations could be automated, the resulting accuracy could not be determined without real-world knowledge of the consolidations created. This fact represents a significant consideration for anyone attempting to replicate this methodology on other data sets.

The sixth consolidation rule relates to the possibility that the first name and last name attributes have been inadvertently switched in the source data sets. Allowing a consolidation where two ERT records match exactly on date of birth and SSN, including an exact match on first-to-last name and last-to-first name provides what is essentially a

sixth build rule utilizing exact matching on four of the view attributes. This rule does not identify many consolidations, reducing the 713,672 consolidated entities of the ERT by 20 to 713,652, however the rule proves more useful in the identity resolution of transaction data sets.

4.2.4 qTR In Use

The seventh consolidation rule requires exact matching on three of the five attributes (date of birth, last name, and SSN) and a qTR value greater than or equal to 0.25 for first name. This is a slight adjustment of consolidation rule one, differing only in the first name comparison requirement. Consolidations identified by this rule will represent students who attended more than one school district with a variation in the first name attribute value which is an approximate (but not exact) string match. These first name variations may be the result of different spellings of the same first name, nicknames, the inclusion of a middle initial, or a combined first name and middle name recorded in the first name attribute. Following the implementation of the seventh consolidation step, the number of student identities changed from 713,652 to 708,102, indicating that a change in the school district and a slight change to the first name occurred for 5,552 records (about 0.6% of the ERT).

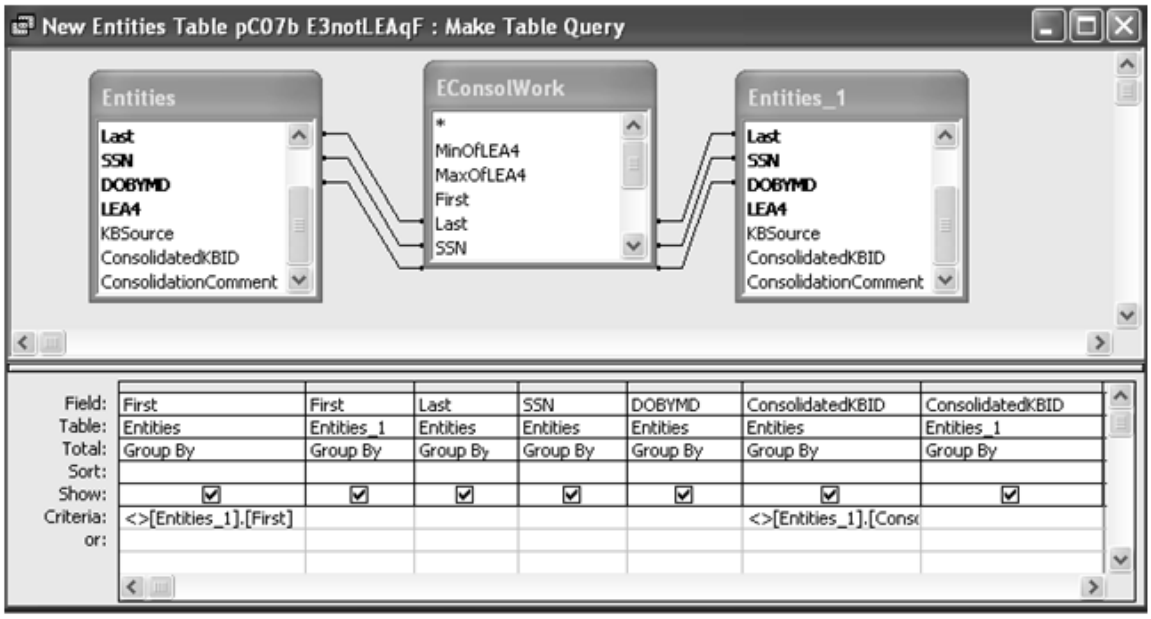

Figure 12. Graphical user interface visualization of the seventh rule of ERT consolidations.

The eighth consolidation rule is related to the seventh rule, this time requiring exact matching on three of the five attributes (date of birth, first name, and SSN) and a qTR value greater than or equal to 0.25 for last name. In addition to being very similar to rule seven, this is a slight adjustment of consolidation rule one, different only in the last name comparison requirement. Consolidations identified by this rule will represent students who attended more than one school district with a variation in the last name attribute value which is an approximate (but not exact) string match. These last name variations may be the result of different spellings of the same last name, a combined last name (with or without a hyphen), or a truncation of the last name. Following the implementation of the eighth consolidation step, the number of student identities changed from 708,102 to 706,177, indicating that a change in the school district and a slight change to the last name occurred for 1,925 records (about 0.2% of the ERT).

4.2.5 Less Confident Consolidations

The ninth rule of consolidation requires exact matching for three of the five attributes (first name, last name, and SSN). These records are indicative of some type of change in the date of birth as recorded by two or more school districts. These changes may occur due to typographical errors or defaulted date values, which show variations across school districts' enrollment data. Following the implementation of the ninth consolidation step, the number of student identities changed from 706,177 to 701,133, indicating that a change of the date of birth value and school district occurred for 5,044 records (about 0.6% of the ERT) for two or more records across school districts.

The tenth rule of consolidation requires exact matching for three of the five attributes (date of birth, last name, and SSN). These records are indicative of some type of change in the first name as recorded by two or more school districts. These first name variations may be the result of different spellings of the same first name, nicknames, the inclusion of a middle initial, or a combined first name and middle name recorded in the first name attribute as recorded across school districts' enrollment data. Following the implementation of the tenth consolidation step, the number of student identities changed from 701,133 to 698,976, indicating that a change of the date of birth value and school district occurred for 2,157 records (about 0.2% of the ERT) for two or more records across school districts.

The eleventh rule of consolidation requires exact matching for three of the five attributes (date of birth, last name, and SSN). These records are indicative of some type of change in the first name as recorded by two or more school districts. These last name variations may be the result of different spellings of the same last name, a combined last

name (with or without a hyphen), or a truncation of the last name as recorded across school districts' enrollment data. Following the implementation of the eleventh consolidation step, the number of student identities changed from 698,976 to 698,092, indicating that a change of the date of birth value and school district occurred for 884 records (about 0.1% of the ERT) for two or more records across school districts.

The twelfth rule of consolidation requires exact matching for three of the five attributes (date of birth, first name, and last name). These records are indicative of some type of change in the SSN as recorded by two or more school districts. These changes may occur due to typographical errors or in cases where one school district has a policy to use some value other than actual SSN as the SSN value. Following the implementation of the twelfth consolidation step, the number of student identities changed from 698,092 to 685,024, indicating that a change of the SSN value and school district occurred for 13,068 records (about 1.5% of the ERT) for two or more records across school districts.

The thirteenth rule of consolidation requires exact matching for two of the five attributes (date of birth and SSN) and requires a qTR value of greater than or equal to 0.25 for a first name match and also for a last name match. These records are indicative of a slight change in both the first name and the last name as recorded by two or more school districts. These name variations may be the result of different spellings of the same names as recorded across school districts' enrollment data. Following the implementation of the thirteenth consolidation step, the number of student identities changed from 685,024 to 684,859, indicating that a change of the first name, last name,

and school district occurred for 165 records (about 0.02% of the ERT) for two or more records across school districts.

The fourteenth rule of consolidation requires exact matching for two of the five attributes (date of birth and SSN) and requires a qTR value of greater than or equal to 0.25 for a first name match. These records are indicative of a slight change in the first name with a differing last name as recorded by two or more school districts. The matching date of birth and SSN values provide strong evidence that the similar first names are describing the same student. The first name variations may be the result of different spellings of the same name as recorded across school districts' enrollment data. Following the implementation of the fourteenth consolidation step, the number of student identities changed from 684,859 to 684,730, indicating that a slight change of the first name, different last name, and different school district occurred for 129 records (about 0.02% of the ERT) for two or more records across school districts.

The fifteenth rule of consolidation requires exact matching for two of the five attributes (date of birth and SSN) and requires a qTR value of greater than or equal to 0.25 for a last name match. These records are indicative of a slight change in the last name with a differing first name as recorded by two or more school districts. The matching date of birth and SSN values provide strong evidence that the similar last names are describing the same student. The last name variations may be the result of different spellings of the same name as recorded across school districts' enrollment data. Following the implementation of the fifteenth consolidation step, the number of student identities changed from 684,730 to 684,660, indicating that a slight change of the first

name, different last name, and different school district occurred for 70 records (about 0.01% of the ERT) for two or more records across school districts.

It is important to note that two records need only satisfy a single rule to instantiate a consolidation. For example, provided two records satisfy the second rule, a consolidation will take place and there is no need to attempt the third, fourth, etc., rules. The application of each rule impacts as many records as necessary during ERT consolidations, however, each rule is implemented independent of the other rules.

4.3 Observations

4.3.1 Unique Name Combinations

Utilizing a number of indications of uniqueness, it is possible to identify additional high-confidence identity consolidations utilizing the qTR methodology with the SSN and date of birth attribute values. Additionally, the frequency of first name and last name values can be incorporated into this evaluation. The sixteenth type of consolidation uncovers a variety of attribute value variations which have prohibited proper identity consolidation in the prior rules. While there are only a small number of identities impacted by this search, it is a good example of the type of consolidation that traditional record matching algorithms will miss. Following the implementation of the sixteenth consolidation step, the number of student identities changed from 684,660 to 684,488, indicating that a slight change of both the SSN and date of birth values occurred for 172 records (about 0.02% of the ERT).

The seventeenth consolidation rule requires exact matching only on the SSN value, but also requires at least two of: qTR >= 0.25 for first name, qTR >= 0.25 for last

name, and qTR >= 0.50 for the tens digit of the year of birth with the month and day of birth. This rule guarantees that in addition to a matching SSN, two additional clues are present in the consolidation rule which should not occur unless the student identity is the same. Following the implementation of the seventeenth consolidation step, the number of student identities changed from 684,488 to 683,924, indicating that a slight change in values for two of the three attributes of first name, last name, and date of birth occurred while SSN was unchanged for 564 records (about 0.06% of the ERT).

Table 13 Consolidated identity counts following implementation of ERT consolidation rules

Rule	DOB	FN	LN	SSN	LEA	Comment	Consolidated Identities
1	X	X	X	X		Change of LEA only	138,736
2	X	X	X		X	Change of SSN only	10,854
3		X	X	X	X	Change of DOB only	4,200
4	X		X	X	X	Change of First name only	6,179
5	X	X		X	X	Change of Last name only	6,139
6	X	LN	FN	X		Reversed First/Last names	20
7	X	q	X	X		Exact D, L, S, qTR First name	5,552
8	X	X	q	X		Exact D, F, S, qTR Last name	1,925
9		X	X	X		Exact F, L, S	5,044
10	X		X	X		Exact D, L, S	2,157
11	X	X		X		Exact D, F, S	884
12	X	X	X			Exact D, F, L	13,068
13	X	q	q	X		Exact D, S, and qTR F, qTR L	165
14	X	q		X		Exact D, S, and qTR F	129
15	X		q	X		Exact D, S, and qTR L	70
16	q	U	U	q		Unique F+L, qTR D, qTR S	172
17	q	q	q	X		Exact SSN, 2 / 3 qTR D,F,L	564

The possibility of false positives, over-consolidations occurring whenever two individuals are inadvertently assigned the same identifier, are a key factor in the determination of which rules to apply. In this research, a set of known examples were

utilized to test the likelihood that a given rule may produce false positives. In the event that a rule showed any likelihood that false positives would occur, the rule was not incorporated into the consolidation process. No false positives are acceptable in the identification of students, and the attributes associated with the type of consolidation utilized are important evidence in the event that a false positive is identified at a future date. The ability to determine the exact nature of the over-consolidation is vital when determining the correct course of action to eliminate the problem from the rule set.

4.3.2 Differences in Concept versus Implementation

Conceptually, it is simple to imagine a scenario where matching records are assigned the same consolidated identifier. As we have seen, the implementation can lead to improper (unsynchronized) identity groupings, which need to be further resolved through a lowest group identifier correction step. Additionally there is a logical complexity which arises given the resolution of two or more ERT records.

Table 14 Example consolidated records of an ERT

Entity ID	Attribute A	Attribute B	Attribute C	Consolidated ID	Consolidation Comment
381	X	23	G	381	Rule 2
437	X	24	G	381	Rule 1, Rule 2
982	Y	24	J	381	Rule 1, lowest

Suppose the next record (1056) added to the ERT is the one shown in Table 15.

Table 15 Additional record (1056) to be incorporated into the example consolidated ERT

Entity ID	Attribute A	Attribute B	Attribute C	Consolidated ID	Consolidation Comment
1056	X	24	J		

While it is true that Entity 1056 share two of the three attribute values with Entities 437 and 982, it is also true that Entity 1056 shares all three of the attribute values with the composite of all ERT records ("intra-group" attribute values) which have been assigned the consolidated ID 381. Should Entity 1056 be noted as a consolidation with 381 because it matches two of the three attributes of a single consolidated 381 record, or should Entity 1056 be noted as a consolidation with 381 because it matches all three "intra-group" attributes values? Consider the difference between Entity 1056 and Entity 191 below.

Table 16 Additional record (191) to be incorporated into the example consolidated ERT

Entity ID	Attribute A	Attribute B	Attribute C	Consolidated ID	Consolidation Comment
191	Z	24	G		

Entity 191 also demonstrates a match with a consolidated ERT record for 381 on two of the three attribute values, however, Attribute A is not equal to any other Attribute A in the example ERT. As a result, Entity 1056 appears to have a different type of affiliation with the consolidated entities of 381 than Entity 191, since 1056 does match all

three “intra-group” attribute values (though not to any single 381 consolidate record). The consolidation comment should reflect the differences in these resolution evidences. A different query methodology is needed to identify these situations, involving multiple self-join statements for the ERT. The numbering utilized in the prior table examples are not indicative of any specific values, only an indication that consolidations may occur anywhere in the ERT, not necessarily in sequential rows or according to any numeric pattern.

Also discovered in the implementation of the consolidation rules are the following scenarios:

- Consider a data set containing only three numbered records. If Rule A links Records 2 and 3 assigning both the consolidated ID “2”, and if Rule B links Records 1 and 3 assigning both the consolidated ID “1”, then Record 2 is no longer part of the consolidation and there are no “clues” to the situation because Record 2 is self-referencing (pointing to the consolidated ID “2”). It appears that Record 2 is unconsolidated, though it should be part of the group.
- Consolidation comments can be duplicated within the string whenever a consolidation adds another record. This situation appears to be related to two consolidated ID values being further consolidated by new consolidation matching, resulting in the consolidation comment append occurring twice.

4.3.3 Nicknames (Frequent Name Pairs)

The consolidated ERT provides an opportunity to create a nickname table specific to the identities which have been consolidated with more than one first name attribute value. Though every consolidation of two or more first names is not indicative of a true

nickname situation, it is possible to sort by the most frequently occurring first name pairs. Implementing this type of evaluation results in 420 first name pairs which occur at least three times (three different student identities) in the ERT. Included among these 420 pairs are cases of first name initials. These do not represent nicknames since it is possible to use only a first initial with every first name, and there is no reciprocal equivalency where a single letter indicates only one first name. Removing any first names of length 1 or 2, results in 406 first name pairs. The table below demonstrates the most frequently occurring first name pairs in the ERT.

Table 17 Most frequently occurring first name pairs in consolidated identity records

First Name 1	First Name 2	Count of Identities
CHRISTOPHE	CHRISTOPHER	241
CHRIS	CHRISTOPHER	141
JOSH	JOSHUA	91
MICHAEL	MICHEAL	88
JONATHAN	JONATHON	83
JEFFERY	JEFFREY	66
ZACHARY	ZACHERY	47
ALEX	ALEXANDER	46
JOHNATHAN	JONATHAN	45
MATHEW	MATTHEW	41
ZACHARY	ZACKARY	38
JORDAN	JORDON	34
NICHOLAS	NICK	31
ZACH	ZACHARY	30
BRITTANY	BRITTNEY	29

It is now possible to utilize these nicknames in future evaluations of either the consolidation rules or the identity resolution of data sets. While the qTR comparison

with a result of greater than or equal to 0.25 has been used to identify many of the first name pairs without the use of a nicknames table, the resulting 406 first name pairs include 58 first name pairs with a qTR value below 0.25. Some of these are traditionally identified as nicknames, such as "TOMMY" to "THOMAS" (qTR = 0.11) and "JACOB" to "JAKE" (qTR = 0.15), while others would be unlikely to be included in traditional nicknames such as "KESHIA" to "KEISHA" (qTR = 0.18) and "KIERRA" to "KERRIA" (qTR = 0.23). This ontological approach reflects not only common nicknames of value to multiple organizations, it also reflects specific value to the Arkansas Department of Education as demonstrated by the data sources utilized to generate this particular set of name pairs (Jamadhvaja & Senivongse, 2005).

It is also possible to utilize the frequencies of consolidation groupings to determine an "anti-nickname" data set. Whenever similar names are shown to occur more frequently among unconsolidated groups, it can be determined that despite the name similarities, no equivalence exists. Examples would be "MARIA" and "MARIO". Though the names differ by only one vowel in the last position, the two names should not be considered equivalent (or nicknames) because of the unlikelihood that an individual with one name would be the same individual as someone with the other name (in this case, due to gender). Because there may be millions of name pair combinations, this type of research may require significant processing time, however, the concept can be shown with a subset of the data. In the table below, a selection of similar names which show no frequency as being nicknames (no examples of being the same individual) which begin with the letter "A" have been displayed as examples.

Table 18 Examples of similar first names which do not occur as aliases (nicknames) in observed consolidated pairs

First Name 1	Count	First Name 2	Count	Consolidated Name Pairs
ADAM	1326	AMY	855	0
ADDISON	254	ALLISON	1155	0
ADRIAN	653	ADRIANA	246	0
ADRIAN	653	ADRIENNE	163	0
ALAN	358	ANNA	1514	0
ALBERT	159	AMBER	2530	0
ALEC	245	ALEXIS	2269	0
ALEJANDRA	208	ALEJANDRO	323	0
ALEX	1365	ALEXA	264	0
ALEXA	264	ALEXANDER	1555	0
ALLEN	415	ALLIE	242	0
AMANDA	2497	ARMANDO	118	0
ANDRE	232	AUDREY	355	0
ANDREA	1024	ANDREW	3221	0
ANDREA	1024	ANDRES	154	0

4.3.4 Implementation Issues

Whenever a new record provides the appropriate attribute values to consolidate either of two existing identities consolidation groups, the SQL statements do not address the full situation of the consolidation possibilities. Suppose there are four records consolidated to ID 89. Suppose there are three records consolidated to ID 105. Now suppose that a new record is introduced to the ERT which indicates that the records consolidated as ID 89 are the same identity as the records consolidated as ID 105. The identity of the new record will be assigned as 89 (even though 105 was equally valid), and the consolidation group identified as 105 will remain unchanged. A special

evaluation is necessary to determine that consolidation group 105 can also be consolidated to ID 89.

One method to accomplish this additional consolidation is to re-run the consolidation rules following any changes in the ERT record composition. As seen in the consolidations to this point, the number of ERT records has remained constant at 879,780. The addition of new records to the ERT will potentially impact the number of consolidations possible for the entire set, not just for the new records added to the ERT. Humans and machines usually reach conclusions on the basis of incomplete knowledge. These conclusions may change when new knowledge becomes available. [McCarthy 2003] Additional research is needed to determine if approaches short of re-running all the consolidation rules are equally valid to address the type of group consolidation mentioned in the 89-to-105 example. It is not the intention of the ERT to determine which of these eight records represents the most accurate (and current) attribute values for the entity (De Amo, S., De Amo, R., Carnielli, & Marcos, 2001, and Greco G., Greco S., & Zumpano, 2003). Because each of these eight records has been recorded in the enrollment source data sets, the knowledge that each represents the same identity is sufficient for proper identification and resolution as needed.

4.4 Transaction Data Set Resolutions

Table 19 Results of the proposed methodology for identity resolution demonstrated with transactional data sets (assessments)

Data Set	Quantity	Identities Resolved	Percent Resolved
ACT (College Board) FY 2007	48,258	47,980	99.4%
ACT (College Board) FY 2008	50,376	50,204	99.7%
ACT (College Board) FY 2009	56,611	56,430	99.7%
Explore 2008	25,119	25,119	100%
Explore 2009	24,447	24,447	100%
Plan 2008	25,442	25,442	100%
Plan 2009	26,016	26,012	99.98%

The unresolved records from the ACT data sets are primarily associated with individuals whose date of birth indicates that they were not in the public school system recently enough to be included in any enrollment data from 2006-2009.

Unlike the consolidation of the ERT, the identity resolution process for transaction files requires only one resolution per record. Assuming the first identity resolution rule succeeds for a particular record, the record is not included in any subsequent rule applications. In practice, the first rule of identity resolution applies to as many records as possible, providing a positive identification of the individuals. Once these identifications have been made, the records are excluded from any attempts to determine identities for the remaining unresolved subset of records.

In summary, the application of the methodology described in this research results in very high identity resolution rates for the transaction data sets tested. In the next section of this research, these results will be compared to the prior methodologies utilized by the organization.

CHAPTER 5

IMPACT OF RESEARCH

5.1 Previous Methodologies

Previous methodologies for identity resolution in the Arkansas Department of Education involved the use of single year enrollment data compared to transaction records via SQL statements which were similar to the first few steps of the consolidation process. Exact matching on the five key attributes (SSN, date of birth, first name, last name LEA) provided approximately the same match rate as seen in the methodology in this research, some attempt was made to resolve the identities of the remaining records via other SQL statements or manual comparison to the same year enrollment data. When exact matching on all five attributes failed, the exact matching on SSN, date of birth, first name, and last name was deemed valid in the event of a student's change in school district. Failing this, a visual inspection was often required to determine the likely causes for unmatched records, with some attempt to systematically identify students either via queries or hand-coding individual records. Variations on the first few consolidation rules utilized by the identity resolution methodology of this research were attempted manually, as determined by the researcher.

Rather than an approximate string matching algorithm, partial matching was allowed using the first few characters of the first name field. The order of these queries eliminated the most complete matches first, relaxing the matching requirements only for those remaining unmatched records. When applicable (such as Explore and Plan assessment data), each record in the same year enrollment file was allowed to match only once to the transaction data. This one-to-one matching eliminated the possibility that a

more relaxed matching rule would incorrectly assign an identifier which had already been used. This process was time-consuming (described by the researcher as approximately one day per transaction file) and resulted in fewer identifications (around 90% to 95%) when compared to the methodology described and implemented in this research. Perhaps most importantly, this process was manual and unlikely to be repeated in the same way on any two occasions. Rather than a strategic, automated, or repeated approach, the researcher relied upon memory of past identity resolutions and any additional skills that had been gained since previous efforts.

Table 20 Results of the prior methodology for identity resolution demonstrated with transactional data sets (assessments)

Data Set	Quantity	Identities Resolved	Percent Resolved
ACT (College Board) FY 2008	50,376	48,705	96.7%
ACT (College Board) FY 2009	56,611	53,235	94.0%
Explore 2008	25,119	23,627	94.1%
Explore 2009	24,447	22,154	90.6%
Plan 2008	25,442	24,274	95.4%
Plan 2009	26,016	23,672	91.0%

Discussion of methodology among other states' departments of education indicated that it is not unusual for the number of identifications in similar transaction data sets to be as low as 85%. The sophistication of many of these states' methodologies is lower than that of Arkansas, as shown by match rates significantly below the amounts achieved by previous Arkansas methodologies. The level of expertise in both the subject matter (education data) and matching methodologies is clearly reflected in the sophistication and results of the identity resolution efforts for the organization. It is understood that several states incorporate third-party identity resolution software in their

systems, however, it is unknown how these vendor products perform in comparison to the research outlined here. Future research into this type of comparison is recommended.

It is known that many other states consider the Social Security Number (SSN) as a primary key for student identifications. As a result, they are susceptible to a number of problems resulting from incorrect SSN values, missing SSN values, and data sources which may utilize some other type of student identifiers. Despite the problematic nature of the SSN, it has remained in use primarily due to either a lack of a viable alternative value or an inability to obtain (or afford) a more comprehensive student identification and resolution system.

5.2 Processing Time

Previous methodologies for identity resolutions within the Arkansas Department of Education were time-consuming. The primary researcher described the process as an "all-day effort", indicating that the bulk of the eight hours of labor were spent writing queries and deciding next steps through visual inspection of both the transaction data sets and the enrollment data being matched. Additionally, these full day efforts did not benefit from any type of automation of the process for future processing. Rather than spending a full day to develop a repeatable methodology which could be automated for quicker future processing, the efforts were a full day as a constant measure of time required and not an initial cost followed by increased production.

Conversely, the automation of the consolidation rules described in this research required less than a day of development. Following this initial time requirement, the processing time for a re-run of the consolidation of the ERT is approximately one hour, and is automated to work without user intervention beyond the initial extract, transform,

and load of the source data sets. Similarly, the time required to develop the identity resolution steps for transaction data sets is only a few hours. The time required to process an average size transaction data set (around 100,000 records) is approximately 15 to 30 minutes, depending upon the number of records which remain unresolved following the first few rules of resolution. This process is also automated to run with minimal user supervision.

Given the increasing number of data requests to the organization, the ability to complete identity resolutions at a higher rate of resolution with significantly lower processing time is a two-fold success. Both goals of this research are achieved for the Arkansas Department of Education, and it is believed that other organizations could also benefit from this type of knowledge-driven identity resolution methodology.

5.3 Vendor Comparison

The Arkansas Department of Education currently allows a third-party vendor to provide identity resolution in the system of record. The vendor's system provides several levels of service, including transcript records and a desktop interface to the data. The identity resolution provided by this vendor has been accepted as part of the larger software service, and to date, no significant research has been conducted to determine the quality of the vendor's identity resolution methodology.

In order to compare the knowledge-driven identity resolution methodology in this research to the proprietary methodology implemented by the vendor, a compilation of the vendor's enrollment files for five previous years was created and included the vendor's individual identifiers as assigned in those years of record. Because the ERT described in this research contains only four years of enrollment data, it is understood that a portion of

the test data set will be unresolved. The test file consists of 722,133 records. It is estimated that approximately 35,646 records (4.94%) will be unresolved, representing students appearing only in the fifth year's data set. This number was estimated utilizing the number of records in the most recent student enrollment file (463,405) divided by 13 (estimating a single grade in K-12).

The identity resolution methodology developed in this research was applied as if this file was a transactional data set. Following the resolution steps for a transactional data set, 37,273 records (5.16%) remained unresolved and would be assigned separate identifiers by the proposed methodology. This is only slightly higher than the 35,646 predicted, a difference of 0.22% of the total data set volume. More important than the number of unresolved records is the number of resolutions which demonstrate the appropriate consolidation of the same entities.

The vendor individual identifiers associated with these 722,133 records represent 624,178 individuals. Whenever the proposed methodology assigns two or more individual identifiers, it can be concluded that a consolidation was missed by the proposed system. Of the 624,178 individuals, the proposed methodology failed to consolidate 451. The other 623,727 appear to have been consolidated correctly, a rate of 99.93%.

The proposed methodology identifiers associated with the 722,133 records represent 615,635 individuals. Whenever the vendor's methodology assigns two or more individual identifiers, it can be concluded that a consolidation was missed by the vendor's system. Of the 615,635 individuals, the vendor's methodology failed to consolidate 17,885. The other 597,750 appear to have been consolidated correctly, a rate of 97.1%.

This rate of consolidation is only slightly higher than that of the previously-utilized sophisticated methodology of the Arkansas Department of Education, and significantly outperforms several states who report normal rates closer to 85%.

Comparing the 99.9% achieved by the methodology proposed in this research to the 97.1% achieved by the third-party vendor's methodology demonstrates the apparent superiority of the proposed methodology in this research to the identity resolution methodology currently being implemented by a third-party vendor for the state of Arkansas. Further study is recommended in this area.

Table 21 Comparison of consolidation results in the proposed methodology and the methodology of the third-party vendor currently in use

Methodology	Test records	Individuals identified	Consolidations missed	Rate of consolidation
This research	722,133	624,178	451	99.93%
Third-party vendor	722,133	615,635	17,885	97.09%

It should be noted that the third-party vendor's 97.1% consolidation rate is a yearly representation of the identity resolution across four years of enrollment data. It is likely that a portion of the 2.9% unconsolidated individuals were actually consolidated in newer data sets. While the vendor may have incorrectly identified an individual in two consecutive years (resulting in an unconsolidated identity in the test data set), it is possible that the vendor's system could correctly resolve the identity in the third and fourth years of data.

It is likely that the vendor's identity resolution problems are resolved over time by the vendor's current system, however, the individual years of data evaluated appear to

have nearly 3% of records unconsolidated. It is unlikely that a single factor causes this problem, rather, it is likely to be a combination of the factors explored in this research, namely expected changes in LEA, unexpected errors of data entry, inconsistent use of SSN, and name preferences or utilization which changes due to the longitudinal aspect of the data. A longer-term study of both systems is recommended before a conclusion is made on this matter.

5.4 Closed-Set Logic

In the special event that a transaction data set is determined to be a closed set, additional rules are possible in identity resolution. The meaning of a closed set is that all of the entities in the transaction data are believed to occur in the ERT, in particular, these records should occur within a single source of the ERT. For instance, a file containing students from the 2008-2009 school year should match directly with the student enrollment data from 2008-2009. The transaction data should not contain any students whose identity is not within the enrollment source. As a result of this situation, identity resolutions may increase in confidence whenever the closed set eliminates possible identities which would otherwise be considered.

An example of this type of increase in confidence would be the case that the ERT contains two records for “DAVID SMITH” at a particular LEA (local education agency). In the event that one of the DAVID SMITH identities corresponds to a student who graduated in 2007, it can be confidently stated that the identity for a DAVID SMITH in 2008-2009 closed-set transaction data would not belong to the 2007 graduate. Without knowing that the transaction data corresponds to the 2008-2009 school year, the identity of DAVID SMITH would be ambiguous for two possible identities.

Additionally, closed set logic may further increase positive identifications in the event that a transaction data set is determined to be both closed and de-duplicated. Whenever a student appears only once in a transaction data set, it becomes possible to eliminate identities from the reference data as they are utilized. As a result, the decreasing number of possible identities in the reference data increases the probability and confidence of identity resolutions for the remaining transaction records. For example, consider a closed set de-duplicated transaction data set which contains two records for "JESSICA JONES". The first of the two records includes a date of birth while the second does not. If the reference data (such as the school year enrollment data) contains two distinct identities for JESSICA JONES, it can be deduced that a match on the single available date of birth value will allow positive identification of both JESSICA JONES identities. The first identity is determined by the matching name and date of birth, while the second identity is determined by the only remaining possibility of identification in the two data sets.

While it is beneficial to determine whenever a transaction data set is believed to be a closed set or a de-duplicated set, it should not be assumed that nothing can go wrong. The dynamic nature of the student population often creates situations within the data which are unanticipated. For instance, the state of Arkansas may require that all students in the third grade take a particular standardized test on a particular date. It seems conclusive that the resulting transaction data from this event will represent a closed set with no duplications. However, the possibility of a test make-up date (for students who were absent on the day of the test) could allow for a student to change schools and to be required by the new school to (re-)take the examination.

A student in the third grade should know that he or she has recently taken a particular examination, however, the student's records will not contain scores for the test (since it has not yet been scored) and it is possible that the receiving school's officials will err on the side of caution and require this new third grade student to take the test (again) on the make-up date. As a result of this double-testing, the student may appear twice in the transaction data, despite the intentions of the state and testing company to eliminate any duplicates. The special cases of closed sets and de-duplicated sets should be approached skeptically and carefully, despite the benefits provided by closed-set logic in identity resolution.

5.4 FERPA-Compliance Planning

The Federal Education Rights and Privacy Act (FERPA) requires that individually-identifiable information is excluded from any use of education records outside of the education agency. As a result FERPA is often seen as a barrier to multi-agency research, however, the regulations are not impassible. The identity resolution methodology described in this research facilitates the proper identification of students despite the differences in the identifier attribute values across longitudinal data sets. Once this identity has been resolved, it is possible to accurately assert shared entities across multiple data sources within the Department of Education and also those outside of the education agency.

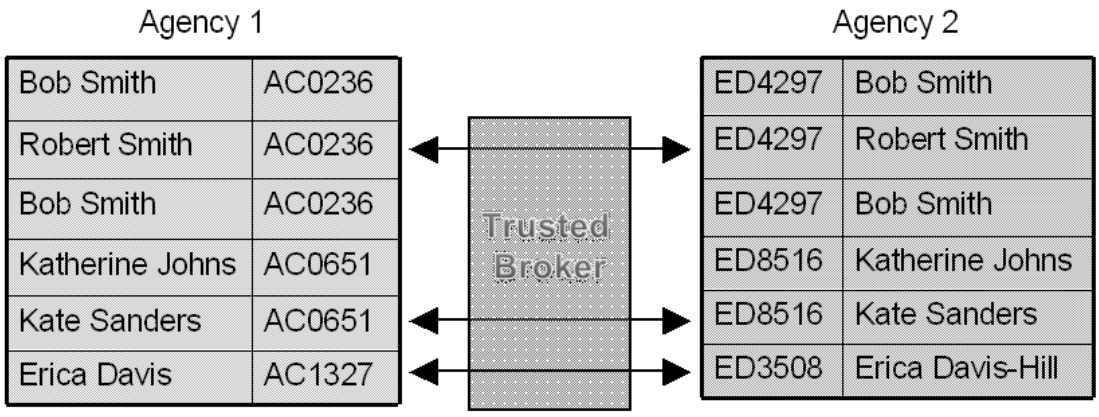

Figure 13. Example of multi-agency identity resolution which provides FERPA-compliant individual identifiers unique to each agency.

Because the identity of the individuals is protected by FERPA, external agencies are required to provide individually-identifying attribute values to the education agency or to some agent acting on behalf of the education agency. These identities provided by the external agencies are traditionally identified by either the Social Security Number (SSN) or some local identifier. It is also possible to determine these identities within the education agency and to provide the requested FERPA-protected data in a non-individually identifiable manner. Of course aggregated data is FERPA-compliant, however, it is not the preferred method of researching student information.

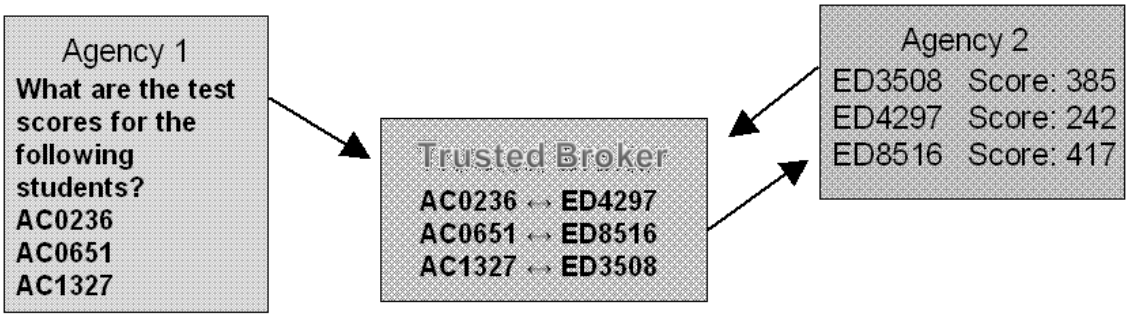

Figure 14. Scenario in which a multi-agency research project remains FERPA-compliant when handled by a trusted broker implementing knowledge-driven identity resolution on behalf of an education agency.

The provision of requested FERPA-protected data in a format which does not allow for the identification of the individual students is also allowable. This can be accomplished by providing the requested data for each individual in no particular order, without any identifying attributes. Any Department of Education is capable of this type of multi-agency data sharing and research, provided they can identify the individuals requested by the external agency and match those individuals to the longitudinal data within the education agency. Once this matching has occurred, the requested education data can be provided to the external agency without identifier attributes, which is allowed by FERPA regulations.

A hybridized version of the trusted-broker's provision of data is also possible in the scenario that the non-education agency provides an additional attribute which identifies groups of individuals. For example, if a non-education agency provided an attribute grouping individuals into groups A, B, C, D, etc., and each of these groups contains more than ten individuals, it would be possible for the education agency to provide the requested education data along with the group attribute value to the non-

education agency. The education data would not be individually identifiable if the group size is at least ten individuals and the resulting data is provided in no particular order for each group.

The agency-specific individual identifiers may be either permanent or temporary, as determined by the trusted-broker according to the likely frequency of repeated usage. In the event that a particular research project is clearly a single request which is unlikely to require any future data from the education agency, the trusted broker may decide to provide individual identifiers which are not retained in the form of a one-to-one mapping with the internal ERT identifier. In other instances, such as the collaboration of the education agency with a non-education agency for longitudinal studies, the trusted broker may retain a detailed record of the one-to-one mappings utilized in the provision of agency-specific identifiers to the non-education agency. In doing so, the trusted broker and the non-education agency may avoid the need to proliferate attribute values such as SSN in future collaborations. In the event that data is lost due to negligence or theft, the agency-specific identifiers provided by the trusted broker system would be of no use in the wrong hands. The same cannot be said if SSN values from any agency are stolen.

This research improves the methodology for identifying individuals within the education agency utilizing the various sources of reference information about the individuals in a longitudinal knowledge-base of resolved identities. Additionally, the use of agency-specific identifiers improves FERPA-compliance in multi-agency research efforts and increases the privacy of individuals by reducing the proliferation of attribute values such as SSN with multiple systems or multiple transactions.

CHAPTER 6

CONCLUSIONS

6.1 Goals Achieved

Three specific goals have been achieved in this research. The first goal achieved is determining a course of action which will facilitate identity-resolved longitudinal education data studies with optimal record linkage for data sets containing varying identifying attributes and attribute values, obtained through various collection methods over a number of years. The planning an implementation of the methodologies in this research address the requirements of this goal.

Secondly, the implementation of a methodology for resolving the representations of real-world entities across multiple longitudinal education data sets which do not utilize a consistent set of identifying attributes has been achieved. The research describes the concepts associated with this type of implementation, and actual longitudinal data sets from the Arkansas Department of Education provided an example of the methodology in practice.

Thirdly, increasing the capability of various state agencies to coordinate research efforts for education data has been achieved by the provision of FERPA-compliant methodology incorporating a trusted-broker concept with the entity reference table (ERT). The reduction in the use of SSN as an identifier is a significant improvement for privacy efforts, and the FERPA-compliant provision of education data enables agencies to coordinate in future research.

In addition to these research goals, two organizational goals were also achieved. First, an increase the quantity of identity resolutions has been demonstrated as compared

to the prior methodology utilized by ADE. Second, the processing time required for identity resolution of transaction data sets has been significantly reduced in the implementation of this research.

6.2 Lessons Learned

Optimization of consolidation and identity resolution rules is heavily dependent upon the implementation of the SQL statement(s). Achieving the desired results of consolidation rules is possible in a variety of methods however the execution time for those queries varies greatly. Research is continually underway to improve the effectiveness of SQL statements and to optimize the required runtimes (Maydanchik, 2007).

Unintended consequences of multi-record consolidations need to be identified and understood. It is possible to overcome these consequences once they have been accurately determined. For example, the consolidation of two previously-consolidated identities often requires a two-step process, rather than a single update statement. The reason for this two-step process is related to the handling of the "linking record" between the two previously-consolidated identities. This "linking record" will be assigned to one of the two consolidated groups, however, the other group will be left unchanged unless the additional knowledge provided by the linking record is utilized appropriately. This utilization for the second group requires a second step in the consolidation process.

There is currently no substitute for the subject matter expertise or "implicit know-how" (Benkler, 2006) of the human data researcher. Although the methodologies described in this research demonstrate significant improvements in the processing time and resolution quantities for the organization, the knowledge provided by experts within

the organization are concomitant with the system's rule sets (Tuomi, 2002). It is unlikely that an organization would be capable of achieving similar levels of success with identity resolution if they are not able to benefit from their own members' varieties of expertise.

6.3 Current and Continual Usage

Currently, the Arkansas Research Center in Conway, Arkansas, handles the majority of data requests in which student identity resolution is necessary for the Arkansas Department of Education. The processes described in this research have been successful in improving both the quantity of resolutions and the speed in which they are obtained. The identity resolution system currently incorporates just over four years of student enrollment data, also incorporating data sets from the student record system managed by a third-party vendor.

In addition to the student identity resolution processes now available, recent work has established a teacher identity resolution system. In the first application of this system, a data set believed to contain recent graduates of Arkansas universities who are predicted to be teaching in the state resulted in just over 50% resolutions. Rather than declare that the teacher identity resolution system has significant flaws, it was determined that the data was actually a superset of the intended group. In addition to those graduates who had become teachers in the state of Arkansas, the data also contained all those who had obtained education degrees but had not been licensed as teachers in the state. The teacher identity resolution system is now available for future reference, consolidations, and resolutions.

Work is presently underway to develop the knowledge-driven identity resolution system for longitudinal education data in parallel on a new platform. Rather than relying

upon a local implementation of the system, a dedicated server is now in place for future development. The process will be migrated to a secured open-source database environment to facilitate increases in utilization and capacity. Over a decade ago, Woodall (1997) noted that "knowledge, culture, and financial investment has made it difficult for organizations to break the old paradigms and move toward the open systems necessary to support today's enterprise." While this fact may persist in many organizations, the Arkansas Department of Education does not share this view. The work presented in this research represents significant progress for knowledge-drive, open-source progress which will save money when compared to existing software investments and move the culture of the department forward in its view of multi-agency research.

6.4 Future Work and Recommendations

The longitudinal nature of the methodology allows researchers to obtain multi-year assessments, progress, growth, and statuses for each resolved identity, facilitating numerous inquiries and evidences for educational research. The Statewide Longitudinal Data Systems (SLDS) Grant Program facilitated through the U.S. Department of Education's Institute for Education Statistics demonstrates the high value placed upon the goals of this research and the research which is facilitated by successful implementation.

The methodologies described in this research demonstrably improve the practices of the Arkansas Department of Education for SLDS programs and provide the capability for FERPA-compliant multi-agency research. The methodologies follow both traditional and innovative approaches to record linkage, utilizing an identity-resolved knowledge-base of reference data. Multiple reference data sets are incorporated into the system in

order to provide all possible views of student identities, as they have been reported in the various systems of record for the Department of Education.

The database structures required by this methodology do not necessarily represent the most optimal configuration possible. Consulting the works of experts such as that of Inmon (1996) would be beneficial to the long-term establishment of the system in a sustainable form. It is unlikely that the ERT, for example, is in a state that would constitute the final word on its own construction or attributes.

Significant progress has been made in the area of total data quality management (TDQM) since the adaptation of earlier works by Ishikawa (1985), Deming (1986), and Juran (1989), to more recent research by Huang (1999) and Lee (2006) conducted through the Massachusetts Institute of Technology's program in Information Quality (MITIQ) established in both industry and academics by Richard Wang, to the current International Association for Information and Data Quality (IAIDQ), among others. Best practices and processes for improving information product and data quality should be incorporated into the methodologies described in this research as well as the organization's data assets as a whole. The support of the organization's management is a key component of successful information quality efforts.

REFERENCES

Benkler, Y., (2006). *The Wealth of Networks*. Oxford Oxfordshire: Oxford University Press.

Bhattacharya, I., Getoor, L. (2006). “A Latent Dirichlet Model for Unsupervised Entity Resolution.” SIAM Conference on Data Mining (SDM).

Bhattacharya, I., Getoor, L. (2007, March). “Collective entity resolution in relational data.” New York: ACM Transactions on Knowledge Discovery from Data, Vol 1, No 1, Article 5.

Bijlsma, M., Koolwaaij, J., Schoneveld, P., Nuijten, J., & Schaafsma, H. (2001). “Semantic reconciliation in supply chain information services.” Enschede: Telematica Instituut.

Borkar, V., Deshmukh, K., Sarawagi, S. (2001, May 21-24). “Automatic segmentation of text into structured records.” Santa Barbara, CA: ACM SIGMOD 2001.

Calvanese, D., De Giacomo, G., Lenzerini, M., Rosati, R. (2004, June). "Logical foundations of peer-to-peer data integration." Paris: Twenty-third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 2004), pp. 241-251.

Christen, P. (2006, September). “A comparison of personal name matching: techniques and practical issues.” Canberra: The Australian National University Technical Reports.

Damerau, F. J. (1964). “A technique for computer detection and correction of spelling errors.” Communications of the ACM, Vol 7, No 3, pp. 171-6.

De Amo, S., De Amo, R., Carnielli, W. A., Marcos, J. (2002). "A logical framework for integrating inconsistent information in multiple databases." International Symposium on Foundations of Information and Knowledge Systems, pp. 67-84.

De Giacomo, G., Lembo, D. (2005). "Data integration with preferences among sources." In ACM IQIS Workshop. pp. 27-34.

Deming, W., (1986). *Out of the Crisis*. Oxford Oxfordshire: Oxford University Press.

Dyché, J., Levy, E., (2006). *Customer Data Integration*. Oxford Oxfordshire: Oxford University Press.

Fellegi, I. P., Sunter, A. B. (1969, December). "A theory for record linkage." Journal of the American Statistical Association, pp. 1183-1210. American Statistical Association.

Gilleland, M. (2009). "Levenshtein distance, in three flavors." Merriam Park. Retrieved October 10, 2009, from http://www.merriampark.com/ld.htm

Gravano, L., Ipeirotis, P. G., Jagadish, H. V., Koudas, N., Muthukrishnan, S., Srivastava, D. (2001). "Approximate string joins in a database (almost) for free." Rome: Proceedings of the 27th VLDB Conference.

Gravano, L. (2001)."Using q-grams in a DBMS for approximate string processing." Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Vol 24, No 4, pp. 28-34.

Greco G., Greco S., Zumpano, E. (2003). "A logical framework for querying and repairing inconsistent databases." IEEE Trans. on Knowledge and Data Engineering, Vol 15, No 6, pp. 1389–1408.

Hall.A., Dowling G.R. (1980, December). “Approximate string matching.” ACM Computing Surveys, Vol 12, No 4, pp. 381–402.

Herzog, T., Scheuren, F., & Winkler, W. E. (2007). *Data Quality and Record Linkage Techniques*. Oxford Oxfordshire: Oxford University Press.

Howe, G. R., Lindsay, J. (1981). “A generalized iterative record linkage computer system for use in medical follow-up studies.” Computers and Biomedical Research 14, pp. 327-340. Academic Press, Inc.

Huang, K., Lee, Y., & Wang, R. (1999). *Quality Information and Knowledge*. Oxford Oxfordshire: Oxford University Press.

Identity resolution. (2010, March 16). In *Wikipedia, the free encyclopedia*. Retrieved March 16, 2010, from http://en.wikipedia.org/wiki/Identity_resolution

Informatica. (2008, May). “Data quality and identity resolution.” (White paper) Redwood City: Informatica Corporation.

Inmon, W., (1996). *Building the Data Warehouse*. Oxford Oxfordshire: Oxford University Press.

Ishikawa, K., (1985). *What Is Total Quality Control? the Japanese Way*. Oxford Oxfordshire: Oxford University Press.

Jamadhvaja, M., Senivongse, T. (2005, November) "An integration of data sources with UML class models based on ontological analysis." Bremen: Proceedings of the First International Workshop on Interoperability of Heterogeneous Information Systems.

Juran, J., (1989). *Juran on Leadership for Quality*. Oxford Oxfordshire: Oxford University Press.

Lee, Y., Pipino, L., Funk, J., & Wang, R. (2006). *Journey to Data Quality*. Oxford Oxfordshire: Oxford University Press.

LingPipe (2006, December). "Code spelunking: Jaro-Winkler string comparison," LingPipe Blog. Retrieved November 22, 2009, from http:\\lingpipe-blog.com/2006/12/13/code-spelunking-jaro-winkler-string-comparison

Maydanchik, A., (2007). *Data Quality Assessment*. Oxford Oxfordshire: Oxford University Press.

McCarthy, J. (2003, January). "Problems and projects in CS for the next 49 years." Journal of the ACM, Vol 50, No 1. pp. 73-79.

Menestrina, D., Benjelloun, O., Garcia-Molina, H. (2006). "Generic entity resolution with data confidences. " Seoul CleanDB

Monge, A. E., Elkan, C. P. (1997). "An efficient domain-independent algorithm for detecting approximately duplicate database records." SIGMOD workshop on data mining and knowledge discovery.

National Archives and Records Administration. (2007, May), "The Soundex indexing system." NARA. Retrieved January 10, 2010, from http:\\www.archives.gov/genealogy/census/soundex.html

Newcombe, H. B., Kennedy, J. M., Axford, S. J., James, A. P. (1959, October 16). "Automatic linkage of vital records." Science, pp. 954-959. American Association for the Advancement of Science.

Newcombe, H. B. (1967, May). "Record linking: The design of efficient systems for linking records into individual and family histories." Chicago: American Journal of Human Genetics, Vol. 19, No. 3.

Newcombe, H. B., Smith, M. E., Howe, G. R., Mingay, J., Strugnell, A., Abbatt, J. D. (1983). "Reliability of computerized versus manual death searches in a study of the health of Eldorado uranium workers." Computers in Biology and Medicine, Vol 13, No 3, pp. 159-167. Pergamon Press Ltd.

Record linkage. (2010, March 16). In *Wikipedia, the free encyclopedia*. Retrieved March 16, 2010, from http://en.wikipedia.org/wiki/Record_linkage

Rud, O., (2001). *Data Mining Cookbook*. Oxford Oxfordshire: Oxford University Press.

Singla, P., Domingos, P. (2006). "Entity resolution with Markov logic." ICDM, Proceedings of the Sixth International Conference on Data Mining, pp. 572-582.

Social Security Administration (SSA). (2010). "Social Security number allocations." SSA. Retrieved March 16, 2010, from http://www.socialsecurity.gov/employer/stateweb.htm

Statistical Society of Canada (SSC). (2008). "Record Linkage." Retrieved March 16, 2010, from http://www.ssc.ca/documents/case_studies/2006/record_index_e.html

Summers, D., (2006). *Quality*. Oxford Oxfordshire: Oxford University Press.

Tachinaba M., Garcia-Molina, H. (2009, January). "Joint entity resolution." Stanford: Technical Report. Stanford InfoLab.

Talburt, J., Wu, N., Pierce, E., Hashemi, R. (2007). "Entity identification using indexed entity catalogs." The 2007 International Conference on Information and Knowledge Engineering, pp. 338-342. Las Vegas, NV: CSREA Press.

Tepping, B. J. (1968, December). “A model for optimum linkage of records.” Journal of the American Statistical Association, pp. 1321-1332. American Statistical Association.

Tuomi, I., (2002). *Networks of Innovation*. Oxford Oxfordshire: Oxford University Press.

Varol, C. (2009, April). “Hybrid matching and risk assessment of the misspelled names.” Little Rock: University of Arkansas at Little Rock Applied Science.

Weisstein, E. W. (2010). "Tetrahedral number." Wolfram – Mathworld. Retrieved February 10, 2010, from http:\\mathworld.wolfram.com/TetrahedralNumber.html

Woodall, J., K., D., & Voehl, F. (1997). *Total Quality in Information Systems and Technology*. Oxford Oxfordshire: Oxford University Press.

APPENDIX A

\\ SQL Statements for the creation and initial population of the Entity Resolution Table (ERT, named here as “Entities”)

```
CREATE TABLE Entities
    (
    KBID        serial DEFAULT GenUniqueID(),
    First       varchar(50),
    Lastvarchar(50),
    SSN         varchar(9),
    DOBYMD varchar(8),
    LEA4        varchar(4),
    KBSource  varchar(255),
    ConsolidatedKBIDint8 DEFAULT 0,
    ConsolidationComment    varchar(255),
    UNIQUE INDEX (KBID),
    PRIMARY KEY (First, Last, SSN, DOBYMD, LEA4)
    );
```

\\ NOTE: GenUniqueID() is a function which generates a random integer value

\\ Initial append to ERT from Student_19 (Student Enrollment 2008-2009)

```
INSERT INTO Entities ( [First], [Last], SSN, DOBYMD, Lea4, KBSource )
SELECT UCase([First]) AS Expr2, UCase([Last]) AS Expr3, Student_19.SSN,
Student_19.DOBYMD, Student_19.Lea4, "S19" AS Expr1
```

FROM Student_19;

\\ Check for existing Student_18 (Student Enrollment 2007-2008) records in ERT

UPDATE Entities INNER JOIN Student_18 ON (Entities.LEA4 = Student_18.Lea4) AND (Entities.DOBYMD = Student_18.DOBYMD) AND (Entities.SSN = Student_18.ssn) AND (Entities.Last = Student_18.lname) AND (Entities.First = Student_18.fname) SET Entities.KBSource = [KBSource] & "S18";

\\ Append Student_18 to ERT (primary key disallows appending of records already existing in the ERT, this prohibition is intentional for this query)

INSERT INTO Entities ([First], [Last], SSN, DOBYMD, Lea4, KBSource)

SELECT UCase([fname]) AS Expr2, UCase([lname]) AS Expr3, Student_18.SSN, Student_18.DOBYMD, Student_18.Lea4, "S18" AS Expr1

FROM Student_18;

\\ This process is repeated for use with Student_17 and Student_16.

APPENDIX B

\\ SQL Statements for the consolidation of the Entity Resolution Table (ERT, named here as “Entities”)

\\ Consolidation 1 - when only LEA differs between ERT records

```
SELECT Min(Entities.LEA4) AS MinOfLEA4, Max(Entities.LEA4) AS
MaxOfLEA4, Entities.First, Entities.Last, Entities.SSN, Entities.DOBYMD,
Min(Entities.KBID) AS MinOfKBID, Count(Entities.KBID) AS CountOfKBID,
Max(Entities.ConsolidatedKBID) AS MaxOfConsolidatedKBID,
Min(Entities.ConsolidatedKBID) AS MinOfConsolidatedKBID INTO EConsolWork
FROM Entities
GROUP BY Entities.First, Entities.Last, Entities.SSN, Entities.DOBYMD
HAVING (((Count(Entities.KBID))>1));


SELECT [Entities].[ConsolidatedKBID],
[EConsolWork].[MinOfConsolidatedKBID] INTO EConsolWork2
FROM Entities INNER JOIN EConsolWork ON
([Entities].[First]=[EConsolWork].[First]) AND
([Entities].[Last]=[EConsolWork].[Last]) AND
([Entities].[SSN]=[EConsolWork].[SSN]) AND
([Entities].[DOBYMD]=[EConsolWork].[DOBYMD])
GROUP BY [Entities].[ConsolidatedKBID],
[EConsolWork].[MinOfConsolidatedKBID];
```

```
UPDATE Entities INNER JOIN EConsolWork2 ON Entities.ConsolidatedKBID
= EConsolWork2.ConsolidatedKBID SET Entities.ConsolidatedKBID =
[MinOfConsolidatedKBID], Entities.ConsolidationComment = [ConsolidationComment]
& "C01";


\\ Consolidation 2 – utilizing exact matching on 4 of the 5 attributes
SELECT Min(Entities.SSN) AS MinOfSSN, Max(Entities.SSN) AS MaxOfSSN,
Entities.First, Entities.Last, Entities.DOBYMD, Entities.LEA4, Min(Entities.KBID) AS
MinOfKBID, Count(Entities.KBID) AS CountOfKBID,
Max(Entities.ConsolidatedKBID) AS MaxOfConsolidatedKBID,
Min(Entities.ConsolidatedKBID) AS MinOfConsolidatedKBID INTO EConsolWork
FROM Entities
GROUP BY Entities.First, Entities.Last, Entities.DOBYMD, Entities.LEA4
HAVING (((Count(Entities.KBID))>1));


SELECT [Entities].[ConsolidatedKBID],
[EConsolWork].[MinOfConsolidatedKBID] INTO EConsolWork2
FROM Entities INNER JOIN EConsolWork ON
([Entities].[DOBYMD]=[EConsolWork].[DOBYMD]) AND
([Entities].[LEA4]=[EConsolWork].[LEA4]) AND
([Entities].[Last]=[EConsolWork].[Last]) AND
([Entities].[First]=[EConsolWork].[First])
```

```
        GROUP BY [Entities].[ConsolidatedKBID],
[EConsolWork].[MinOfConsolidatedKBID];

        SELECT EConsolWork2.ConsolidatedKBID AS ConsKBID,
Min(EConsolWork2_1.MinOfConsolidatedKBID) AS MinOfMinOfConsolidatedKBID
INTO EConsolWork3
        FROM EConsolWork2 INNER JOIN EConsolWork2 AS EConsolWork2_1 ON
EConsolWork2.MinOfConsolidatedKBID = EConsolWork2_1.ConsolidatedKBID
        GROUP BY EConsolWork2.ConsolidatedKBID;

        UPDATE Entities INNER JOIN EConsolWork3 ON Entities.ConsolidatedKBID
= EConsolWork3.ConsKBID SET Entities.ConsolidatedKBID =
[MinOfMinOfConsolidatedKBID], Entities.ConsolidationComment =
[ConsolidationComment] & "C02";

        \\ Consolidation 7 – utilizing the qTR function
        SELECT Min(Entities.First) AS MinOfFirst, Max(Entities.First) AS MaxOfFirst,
Entities.Last, Entities.SSN, Entities.DOBYMD INTO EConsolWork
        FROM Entities
        GROUP BY Entities.Last, Entities.SSN, Entities.DOBYMD
        HAVING (((Max(Entities.First))<>Min([First])));
```

```
SELECT Entities.First, Entities_1.First, Entities.Last, Entities.SSN,
Entities.DOBYMD, Entities.ConsolidatedKBID, Entities_1.ConsolidatedKBID INTO
EConsolWork2
FROM (Entities INNER JOIN EConsolWork ON (Entities.Last =
EConsolWork.Last) AND (Entities.SSN = EConsolWork.SSN) AND
(Entities.DOBYMD = EConsolWork.DOBYMD)) INNER JOIN Entities AS Entities_1
ON (EConsolWork.Last = Entities_1.Last) AND (EConsolWork.DOBYMD =
Entities_1.DOBYMD) AND (EConsolWork.SSN = Entities_1.SSN)
GROUP BY Entities.First, Entities_1.First, Entities.Last, Entities.SSN,
Entities.DOBYMD, Entities.ConsolidatedKBID, Entities_1.ConsolidatedKBID
HAVING (((Entities.First)<>[Entities_1].[First]) AND
((Entities.ConsolidatedKBID)<>[Entities_1].[ConsolidatedKBID]));

SELECT EConsolWork2.Entities_ConsolidatedKBID AS C1,
EConsolWork2.Entities_1_ConsolidatedKBID AS C2 INTO EConsolWork3
FROM EConsolWork2
WHERE (((qTR([Entities_First],[Entities_1_First]))>=0.25))
GROUP BY EConsolWork2.Entities_ConsolidatedKBID,
EConsolWork2.Entities_1_ConsolidatedKBID;

INSERT INTO EConsolWork3 ( C1, C2 )
SELECT EConsolWork2.Entities_1_ConsolidatedKBID,
EConsolWork2.Entities_ConsolidatedKBID
```

```
FROM EConsolWork2
WHERE (((qTR([Entities_First],[Entities_1_First]))>=0.25))
GROUP BY EConsolWork2.Entities_1_ConsolidatedKBID,
EConsolWork2.Entities_ConsolidatedKBID;

SELECT EConsolWork3.C1 AS ConsKBID, Min(IIf([C1]<[C2],[C1],[C2])) AS
MinOfConsKBID INTO EConsolWork4
FROM EConsolWork3
GROUP BY EConsolWork3.C1
ORDER BY Min(IIf([C1]<[C2],[C1],[C2]));

UPDATE Entities INNER JOIN EConsolWork4 ON Entities.ConsolidatedKBID
= EConsolWork4.ConsKBID SET Entities.ConsolidatedKBID = [MinOfConsKBID],
Entities.ConsolidationComment = [ConsolidationComment] & "C07";

\\ Consolidation 16 – utilizing unique first and last name combinations
SELECT [First Last YOB YearCount].First, [First Last YOB YearCount].Last,
[First Last YOB YearCount].CountOfYOB AS Years, [First Last YOB
YearCount].Records INTO EConsolWork
FROM [First Last YOB YearCount] INNER JOIN Entities ON ([First Last YOB
YearCount].Last = Entities.Last) AND ([First Last YOB YearCount].First =
Entities.First)
```

```
GROUP BY [First Last YOB YearCount].First, [First Last YOB
YearCount].Last, [First Last YOB YearCount].CountOfYOB, [First Last YOB
YearCount].Records
HAVING ((([First Last YOB YearCount].CountOfYOB)=1) AND (([First Last
YOB YearCount].Records)>1) AND
((Max(Entities.ConsolidatedKBID))>Min([ConsolidatedKBID])));

SELECT Entities.KBID, Entities.First, Entities.Last, Entities.SSN,
Entities.DOBYMD, Entities.LEA4, Entities.KBSource, Entities.ConsolidatedKBID,
Entities.ConsolidationComment INTO EConsolWork2
FROM Entities INNER JOIN EConsolWork ON (Entities.First =
EConsolWork.First) AND (Entities.Last = EConsolWork.Last)
ORDER BY Entities.Last, Entities.First;

SELECT EConsolWork2.First, EConsolWork2.Last, EConsolWork2.SSN,
EConsolWork2_1.SSN, EConsolWork2.DOBYMD, EConsolWork2_1.DOBYMD,
Min(EConsolWork2.LEA4) AS MinOfLEA4, Max(EConsolWork2_1.LEA4) AS
MaxOfLEA4, EConsolWork2.ConsolidatedKBID, EConsolWork2_1.ConsolidatedKBID,
qTR(Right([EConsolWork2].[SSN],7),Right(EConsolWork2_1.SSN,7)) AS qTR_S,
qTR(Right([EConsolWork2].[DOBYMD],4),Right(EConsolWork2_1.DOBYMD,4)) AS
qTR_D INTO EConsolWork3
```

FROM EConsolWork2 INNER JOIN EConsolWork2 AS EConsolWork2_1 ON (EConsolWork2.Last = EConsolWork2_1.Last) AND (EConsolWork2.First = EConsolWork2_1.First)

GROUP BY EConsolWork2.First, EConsolWork2.Last, EConsolWork2.SSN, EConsolWork2_1.SSN, EConsolWork2.DOBYMD, EConsolWork2_1.DOBYMD, EConsolWork2.ConsolidatedKBID, EConsolWork2_1.ConsolidatedKBID, qTR(Right([EConsolWork2].[SSN],7),Right(EConsolWork2_1.SSN,7)), qTR(Right([EConsolWork2].[DOBYMD],4),Right(EConsolWork2_1.DOBYMD,4))

HAVING (((EConsolWork2.ConsolidatedKBID)<[EConsolWork2_1].[ConsolidatedKBID]) AND ((qTR(Right([EConsolWork2].[SSN],7),Right([EConsolWork2_1].[SSN],7)))>=0.1) AND ((qTR(Right([EConsolWork2].[DOBYMD],4),Right([EConsolWork2_1].[DOBYMD],4)))>=0.1));

SELECT EConsolWork3.First, [First Name Counts].FCount, [Last Name Counts].LCount, EConsolWork3.Last, EConsolWork3.EConsolWork2_SSN, EConsolWork3.EConsolWork2_1_SSN, EConsolWork3.EConsolWork2_DOBYMD, EConsolWork3.EConsolWork2_1_DOBYMD, EConsolWork3.MinOfLEA4, EConsolWork3.MaxOfLEA4, EConsolWork3.EConsolWork2_ConsolidatedKBID, EConsolWork3.EConsolWork2_1_ConsolidatedKBID, EConsolWork3.qTR_S, EConsolWork3.qTR_D, CDbl([qTR_S])+CDbl([qTR_D]) AS Expr1, [FCount]*[LCount] AS Expr2, CLng([qTR_S]) AS Expr3 INTO EConsolWork4

```
        FROM (EConsolWork3 INNER JOIN [First Name Counts] ON
EConsolWork3.First = [First Name Counts].First) INNER JOIN [Last Name Counts] ON
EConsolWork3.Last = [Last Name Counts].Last
        WHERE (((CDbl([qTR_S])+CDbl([qTR_D]))>=0.4) AND
(([FCount]*[LCount])<50000)) OR (((CLng([qTR_S]))>=0.4)) OR ((([First Name
Counts].FCount)<2000) AND (([Last Name Counts].LCount)<20) AND
((CDbl([qTR_S])+CDbl([qTR_D]))>=0.3)) OR ((([First Name Counts].FCount)<20)
AND (([Last Name Counts].LCount)<2000) AND
((CDbl([qTR_S])+CDbl([qTR_D]))>=0.3))
        ORDER BY CDbl([qTR_S])+CDbl([qTR_D]) DESC;


        SELECT Entities.ConsolidatedKBID,
Min(IIf([EConsolWork2_ConsolidatedKBID]<[EConsolWork2_1_ConsolidatedKBID],[
EConsolWork2_ConsolidatedKBID],[EConsolWork2_1_ConsolidatedKBID])) AS
MinOfConsKBID INTO EConsolWork3
        FROM Entities INNER JOIN EConsolWork4 ON (Entities.First =
EConsolWork4.First) AND (Entities.Last = EConsolWork4.Last)
        GROUP BY Entities.ConsolidatedKBID
        ORDER BY
Min(IIf([EConsolWork2_ConsolidatedKBID]<[EConsolWork2_1_ConsolidatedKBID],[
EConsolWork2_ConsolidatedKBID],[EConsolWork2_1_ConsolidatedKBID]));
```

```
UPDATE Entities INNER JOIN EConsolWork3 ON Entities.ConsolidatedKBID
= EConsolWork3.ConsolidatedKBID SET Entities.ConsolidatedKBID =
[MinOfConsKBID], Entities.ConsolidationComment = [ConsolidationComment] &
"C16";


\\ Consolidation 17 – exact SSN and 2 out of 3 qTR on FLD
SELECT Entities.First, Entities_1.First, Entities.Last, Entities_1.Last,
Entities.SSN, Entities.DOBYMD, Entities_1.DOBYMD, Entities.ConsolidatedKBID,
Entities_1.ConsolidatedKBID INTO EConsolWork
FROM Entities INNER JOIN Entities AS Entities_1 ON Entities.SSN =
Entities_1.SSN
GROUP BY Entities.First, Entities_1.First, Entities.Last, Entities_1.Last,
Entities.SSN, Entities.DOBYMD, Entities_1.DOBYMD, Entities.ConsolidatedKBID,
Entities_1.ConsolidatedKBID
HAVING (((Entities_1.ConsolidatedKBID)<[Entities].[ConsolidatedKBID]));


SELECT EConsolWork.Entities_First, EConsolWork.Entities_1_First,
EConsolWork.Entities_Last, EConsolWork.Entities_1_Last, EConsolWork.SSN,
EConsolWork.Entities_DOBYMD, EConsolWork.Entities_1_DOBYMD,
EConsolWork.Entities_ConsolidatedKBID,
EConsolWork.Entities_1_ConsolidatedKBID, qTR([Entities_First],[Entities_1_First])
AS qTR_F, qTR([Entities_Last],[Entities_1_Last]) AS qTR_L,
```

```
qTR(Right([Entities_DOBYMD],5),Right([Entities_1_DOBYMD],5)) AS qTR_D INTO
EConsolWork2
        FROM EConsolWork
        WHERE (((qTR([Entities_First],[Entities_1_First]))>=0.25) AND
((qTR([Entities_Last],[Entities_1_Last]))>=0.25)) OR
(((qTR([Entities_First],[Entities_1_First]))>=0.25) AND
((qTR(Right([Entities_DOBYMD],5),Right([Entities_1_DOBYMD],5)))>=0.25)) OR
(((qTR([Entities_Last],[Entities_1_Last]))>=0.25) AND
((qTR(Right([Entities_DOBYMD],5),Right([Entities_1_DOBYMD],5)))>=0.5));

        SELECT [EConsolWork2].Entities_ConsolidatedKBID AS ConsKBID,
Min([EConsolWork2].Entities_1_ConsolidatedKBID) AS MinOfConsKBID INTO
EConsolWork3
        FROM EConsolWork2
        GROUP BY [EConsolWork2].Entities_ConsolidatedKBID;

        UPDATE Entities INNER JOIN EConsolWork3 ON Entities.ConsolidatedKBID
= EConsolWork3.ConsKBID SET Entities.ConsolidatedKBID = [MinOfConsKBID],
Entities.ConsolidationComment = [ConsolidationComment] & "C17";
```