# Minnesota P-20 Statewide
# Longitudinal Education Data System Charter

## A. Vision

Minnesota will develop a P-20 Statewide Longitudinal Education Data System (SLEDS) to provide educators and policymakers with more comprehensive data and analysis from which to make informed decisions leading to educational improvement at all levels. As required by the federal government as a condition of receiving federal fiscal stabilization funds, this statewide longitudinal data system will match student data from pre-kindergarten through completion of postsecondary, enabling educators and policymakers to answer a range of program and policy questions to gauge the effectiveness of programs and design targeted improvement strategies.

## B. Rationale

State governments across the country are putting plans into place for developing longitudinal data systems for two primary reasons: to facilitate data-driven decision making and because the federal government requires states to complete such systems as a condition of receiving fiscal stabilization and major federal grant funding. In 2008, 28 states reported their ability to match student records between the P-12 and postsecondary systems, according to the Data Quality Campaign. States are in various states of implementing a completed fully functional longitudinal data system.

## C. Legal Authorization

**Minnesota Law**

In the 2008 Minnesota legislative session lawmakers passed statutory language allowing the Minnesota Department of Education and the Minnesota Office of Higher Education to share data elements each currently collects for purposes of conducting research to answer questions identified in the vision for the Statewide Longitudinal Education Data System.

Chapter 298: Sec.2. *M.S*. 2006, section 13.32 Subd.11. was amended to provide for:

> Data Sharing; improving instruction. The following educational data may be shared between the Department of Education and the Minnesota Office of Higher Education as authorized by the Code of Federal Regulations, title 34, section 99.31 (a)(6), to analyze instruction in school districts for purposes of improvement:
>
>> (1) attendance data, including name of school or institution, school district, year or team of attendance, and term type:
>> (2) student demographic and enrollment data;
>> (3) academic performance and testing data; and
>> (4) special academic services received by a student.
>
> Any analysis of or report on the data must contain only summary data.

**Minnesota Interagency Agreements**

Any usage of the P-20 data must adhere to the legal requirements of the following data sharing agreements:

- "State of Minnesota Interagency Data Sharing Agreement" signed by the Minnesota Department of Education and the Minnesota Office of Higher Education on September 1, 2009; and
- "Enrollment Data Sharing Agreement" between the Minnesota Office of Higher Education and higher education institutions providing student enrollment data.

**Federal Laws**

The federal mandate regarding state longitudinal data systems are contained in two federal laws:

- American Recovery and Reinvestment Act
- America Competes Act

The **American Recovery and Reinvestment Act of 2009** 26 United States Code Section 1 is in Title XIV - State Fiscal Stabilization Fund, Section 14006 State Applications. It states:

(d) Assurances.--An application under subsection (b) shall include the following assurances:

(3) Improving collection and use of data.--The State will establish a longitudinal data system that includes the elements described in section 6401(e) (2) (D) of the America COMPETES Act (20 U.S.C. 9871).

The **America Competes Act** lays out requirements for state longitudinal data systems between K-12 and postsecondary education. America Competes Act, 20 United States Code Section 9871. It is in Title 20 – Education, Chapter 78 – Science, Technology, Engineering, Mathematics and Critical Foreign Language Education, Subchapter IV – Alignment of Education Programs. Section 6401 (e) (2) (D):

(D) Required elements of a statewide p-16 education data system.--The State shall ensure that the statewide P-16 education data system includes the following elements:

(i) Preschool through grade 12 education and postsecondary education.--With respect to preschool through grade 12 education and postsecondary education—

(I) a unique statewide student identifier that does not permit a student to be individually identified by users of the system;

(II) student-level enrollment, demographic, and program participation information;

(III) student-level information about the points at which students exit, transfer in, transfer out, drop out, or complete P-16 education programs;

(IV) the capacity to communicate with higher education data systems; and

(V) a State data audit system assessing data quality, validity, and reliability.

(ii) Preschool through grade 12 education.--With respect to preschool through grade 12 education—

(I) yearly test records of individual students with respect to assessments under section 1111(b) of the Elementary and Secondary Education Act of 1965 (20 U.S.C. 6311(b));

(II) information on students not tested by grade and subject;

(III) a teacher identifier system with the ability to match teachers to students;

(IV) student-level transcript information, including information on courses completed and grades earned; and

(V) student-level college readiness test scores.

(iii) Postsecondary education.--With respect to postsecondary education, data that provide—

(I) information regarding the extent to which students transition successfully from secondary school to postsecondary education, including whether students enroll in remedial coursework; and

(II) other information determined necessary to address alignment and adequate preparation for success in postsecondary education.

(E) Functions of the statewide p-16 education data system.--In implementing the statewide P-16 education data system, the State shall

(i) identify factors that correlate to students' ability to successfully engage in and complete postsecondary-level general education coursework without the need for prior developmental coursework;

(ii) identify factors to increase the percentage of low-income and minority students who are academically prepared to enter and successfully complete postsecondary-level general education coursework; and

(iii) use the data in the system to otherwise inform education policy and practice in order to better align State academic content standards, and curricula, with the demands of postsecondary education, the 21st century workforce, and the Armed Forces.


## D. Purpose

The general purpose of the SLEDS system is to identify the predictors of long-term prekindergarten through higher education student success – in other words, define "what makes a difference" in the academic experiences of students. The intention is to link data on students who graduate from a Minnesota public high school and attend a Minnesota post-secondary institution at the undergraduate level (approximately 40,000 high school graduates annually).

Students who might not be captured in the linkage would include:

- Those who graduate from a private Minnesota school (approximately 7 percent of Minnesota high school graduates annually or 5,000 students).
- Those who go to college from a home school environment (approximately 1 percent or 400 students).
- Those who did not graduate from high school but subsequently attend a postsecondary institution (approximately 2 percent or 1,300 students).
- Students who receive a GED (approximately 1 percent or 400 students).
- Students attending a Minnesota postsecondary institution who graduated from an out of state high school.

# E. Research

Three areas of research and analysis can be conducted upon successful construction of the SLEDS system. SLEDS data will provide a comprehensive foundation for documenting the performance of students, schools, and colleges, while improving the ability to address questions about Minnesota's investment in education. Data alone cannot improve performance but it can support the careful consideration of issues and analysis leading to action.

The explanation of areas for research and analysis include example questions that can be examined using SLEDS data. The actual research and analysis to be completed shall be identified and managed per the SLEDS governance structure to be discussed in section G of this charter.

**Area one: System Performance Analysis**

This research area focuses on the performance of the overall educational system, identifying aggregate performance at key points in time (e.g. high school completion, postsecondary participation). These data may also be used to focus on student performance in relationship to criteria established by Minnesota and provide a common rubric for evaluating student and system performance.

Example questions:

- Who participates in higher education upon high school graduation? Who does not participate? What are the characteristics of non-participants?
- What are the patterns of enrollment for Minnesota students? At what rate do students who leave postsecondary education (dropout, withdraw) reenroll?
- What are the patterns of course completion for persisting students? What are the characteristics of persisters?
- What are the characteristics of students who pursue programs in science, technology, engineering and mathematics (STEM) in college?
- What percentage of students in high school technical preparation programs go to college? What types of college do they attend? What programs do they pursue?
- By district, what percent of students are enrolling in, persisting in and completing postsecondary education? Which students require remedial education? What kinds of programs, majors and degrees do each district's students enroll in?
- What is the correlation between K-12 academic tests (MCA scores, ACT scores) and college enrollment, persistence and completion?

**Area two: Educational Attainment Gap Analysis**

This research area focuses on the performance of students defined by their demographic, socioeconomic or geographic characteristics. While certain educational transition points for key groups (e.g. students of color) has been analyzed, longitudinal information is currently not available to identify how performance lags early in the pipeline (e.g. failure to graduate from high school) impact later measures of educational success.

Example questions:

- What are the enrollment, retention, and completion rates for key racial/ethnic groups? Key geographic groups? By gender? By socio-economic status? For students receiving limited English proficiency services in K-12?
- What percentage of high school graduates attend college based on race/ethnicity, high school, school district and socio-economic status? What type of colleges do these high school students attend? At what rates do these students persist in college? What type of postsecondary programs do these students pursue?

**Area three: Program and Intervention Analysis**

This research area focuses on evaluation of educational programs and interventions designed to increase educational attainment. A number of large and small scale intervention programs exist to promote equality of educational outcomes. Some programs have sought to raise the academic achievements and educational aspirations of selected students from lower socioeconomic backgrounds and increase the numbers of these students graduating from high school, enrolling in college, and graduating from college. Although these intervention efforts have been in operation for many years, little is known about their collective impact on the student population. State level financing of such programs would benefit from targeted data on program participants and their educational outcomes compared to peer group performance.

Example questions:

- What courses, curriculum and programs lead to college participation and completion? Are there particular courses (e.g. calculus) or academic pathways (e.g. dual enrollment) that are correlated with academic success in college?
- What are the high school course taking patterns of students enrolling in remedial coursework upon entry into postsecondary education?
- At what rates do students completing dual enrollment programs or courses (e.g. Post Secondary Enrollment Options program, Advanced Placement, International Baccalaureate, College in the Schools) persist in and complete higher education? Do these rates vary by program?
- What percent of GEAR Up participants enroll in higher education? What factors are correlated to postsecondary enrollment for these students?

- Is there a correlation between participation in college access programs while in high school and college participation, persistence and completion? Which programs have the highest enrollment rates? Persistence rates? Completion rates?
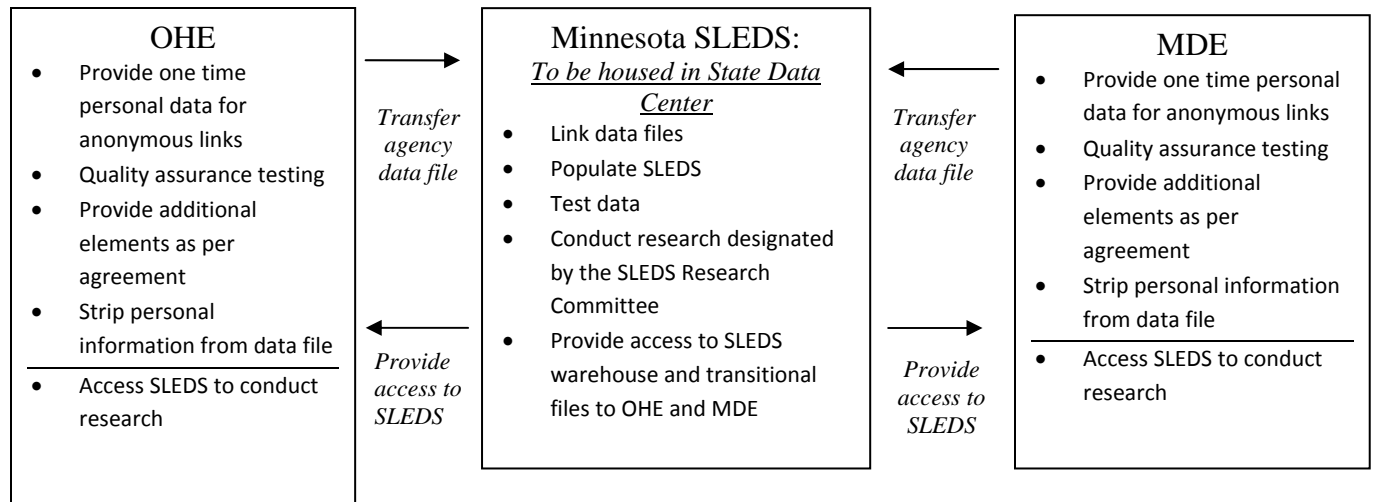
# F. Data Exchange Process

The following process for matching data from the Minnesota Department of Education and the Minnesota Office of Higher Education has been developed. The data matching process maximizes data security, complies with federal and state regulation, and builds a system that can accomplish the intended research outcomes.

**Construction of Data Files**

A.  **Responsible person:** Both MDE and OHE will appoint SLEDS IT staff with access to full student record information from each agency to construct:

1)  a reference table of personally identifiable data used only in the process to match MDE students with OHE students, and

2)  data files used to populate the SLEDS System limited to the variables identified by the SLEDS Research and Data Advisory Committees as approved by the SLEDS Governance Committee.

B.  **Matching Student Records:** Respective SLEDS IT staff produce a reference table used to match students between OHE and MDE. These reference tables will be exchanged between MDE and OHE using existing secure file transfer protocols and technology. The reference table will include personally identifiable information that is stored in both agencies (e.g. first name, last name, date of birth, year of high school graduation, high school, and MARSS number).

   Probabilistic matching methodology will be used to determine how many and which of the student records in both reference tables match exactly and the likelihood that the two records are the same student.

C.  **Assign Anonymous ID:** Each student record in the reference table will be assigned a random anonymous identification number referred to as the "SLEDS Number". The SLEDS number shall be sent to the originating agency to be used to populate the SLEDS data files. After the match process is complete the linked reference table data including all personally identifiable data shall be destroyed.

   Personally identifiable data other than the SLEDS anonymous student identification number shall never be transferred to the SLEDS system.

D.  **Timeframe:** Data exchange shall occur at a minimum yearly – upon the conclusion of each academic year/fiscal year once data sets are finalized.

**Data Exchange Process Visual**

| OHE | | Minnesota SLEDS: *To be housed in State Data Center* | | MDE |
|---|---|---|---|---|
| • Provide one time personal data for anonymous links<br>• Quality assurance testing<br>• Provide additional elements as per agreement<br>• Strip personal information from data file<br><br>• Access SLEDS to conduct research | *Transfer agency data file* →<br><br>← *Provide access to SLEDS* | • Link data files<br>• Populate SLEDS<br>• Test data<br>• Conduct research designated by the SLEDS Research Committee<br>• Provide access to SLEDS warehouse and transitional files to OHE and MDE | ← *Transfer agency data file*<br><br>*Provide access to SLEDS* → | • Provide one time personal data for anonymous links<br>• Quality assurance testing<br>• Provide additional elements as per agreement<br>• Strip personal information from data file<br><br>• Access SLEDS to conduct research |

# G. Data Governance Structure

The SLEDS system will be jointly managed by the Minnesota Office of Higher Education and the Minnesota Department of Education. Federal funding will be sought to hire a project manager. Federal grant funding will also be sought to build the capacity at both agencies to effectively manage the data system.

The following governance structure was adopted by the Minnesota P-20 Education Partnership on April 14, 2010.

*Governance Committee:* The SLEDS Governance Committee standing membership shall include one standing representative from each of MDE, OHE, the Department of Employment and Economic Development  ("DEED"), the Minnesota Association of School Administrators ("MASA"), the Minnesota School Boards Association ("MSBA"), Minnesota Association of Elementary School Principals ("MAESP"), Minnesota Association of Secondary School Principals ("MASSP"), Education Minnesota ("EdMn"), the University of Minnesota, the Minnesota State Colleges and Universities ("MnSCU"), the Minnesota Private College Council ("MPCC"), the Minnesota Career College Association ("MCCA"), and the business community. These members shall be elected and/or appointed by their respective organizations and shall retain membership until their successor is elected and/or appointed.

Two additional members from the Partnership may be appointed to serve a two-year term. These members may not be representatives of the institutions represented by the SLEDS standing membership of the SLEDS Governance Committee.

Each member may have a designee with the authority to represent and make decisions for that member.

The duties of the SLEDS Governance Committee members shall include:

1. Evaluate and consider SLEDS research and evaluation topics at the request of the Minnesota P-20 Education Partnership or other groups that will inform data driven decisions and educational policy formation.
2. Approve requests for additional data elements to be added to SLEDS for research purposes.
3. Sign off on data security protocols and data transfer procedures.
4. Appoint members to the Research Committee and Data Advisory Committee of the SLEDS Governance Committee.
5. Requests for accessing data for research purposes shall be directed to the SLEDS Governance Committee.
6. Ensure access to public data in accordance with state and federal privacy laws.

***SLEDS Research Committee:*** The Research Committee membership shall include one representative each from MDE, OHE, DEED, either MAESP or MASSP, the University of Minnesota, MnSCU, MPCC, MCCA, and two K-12 representatives. The Research Committee duties shall include:

1. Advise in the development of a data system interface/portal to provide access to SLEDS data while adhering to state and federal laws.
2. Review proposals for research and evaluation and make recommendations for implementation.
3. Develop research and evaluation proposals for utilizing the SLEDS data to further state research goals set by the Governance Committee.
4. Provide technical expertise and consultation on research methodologies.
5. Develop protocols for maximizing validity and reliability of SLEDS data.
6. Recommend protocols for allowing non-MDE and non-OHE staff access to SLEDS data.
7. Review data requirements for research and evaluation proposals and recommend additional elements to be collected if necessary to meet research requests.

***SLEDS Data Advisory Committee:*** The Data Advisory Committee members shall include persons with technical expertise about data collection, data structure, data security and protocols for maximizing validity and reliability of SLEDS data. The Data Advisory Committee membership shall include one representative each from MDE, OHE, DEED, the Office of Enterprise Technology ("OET"), the University of Minnesota, MnSCU, MPCC and MCCA, and two K-12 representatives.  The SLEDS Data Advisory Committee duties include:

1. Evaluate research proposals with respect to data-based components and make recommendations to the Research Committee for approval.
2. Provide the technical expertise for the creation of an interface for users that is not only secure but also user friendly and widely available to authorized users at all levels of the educational system.
3. Provide technical expertise and consultation on data structure and data linkages for existing and new data.
4. Develop protocols for maximizing validity and reliability of SLEDS data.

***SLEDS System Coordinators****:* appointed by OHE and MDE shall report to the SLEDS Governing Committee on the research, evaluation and technical operations of the SLEDS.  The Coordinators will:

1. Work with Governance Committee and Agencies to identify funding opportunities to support SLEDS work.
2. Work with IT staff on data security, data transfer and data file construction issues.
3. Respond to data requests and maintain access portal to data.
4. Maintain awareness and compliance with FERPA and other relevant laws.
5. Work with the Data Advisory Committee to coordinate data management including: set data standards, define data elements, document data processes, and identify file specifications.
6. Work with the Research Committee to facilitate research using SLEDS data.

7. Coordinate the SLEDS Research and SLEDS Data Advisory Committees.
8. Represent Minnesota at national conferences related to P-20 systems and research.
9. Serve as spokespersons for the SLEDS system.
10. Assist agencies in public relations aspects of SLEDS in communication with school districts and charter schools, higher education institutions and the public.

The SLEDS Governance Committee shall meet at least twice each year, and the Minnesota P-20 Education Partnership shall review the annual strategic work plan of the Committee.
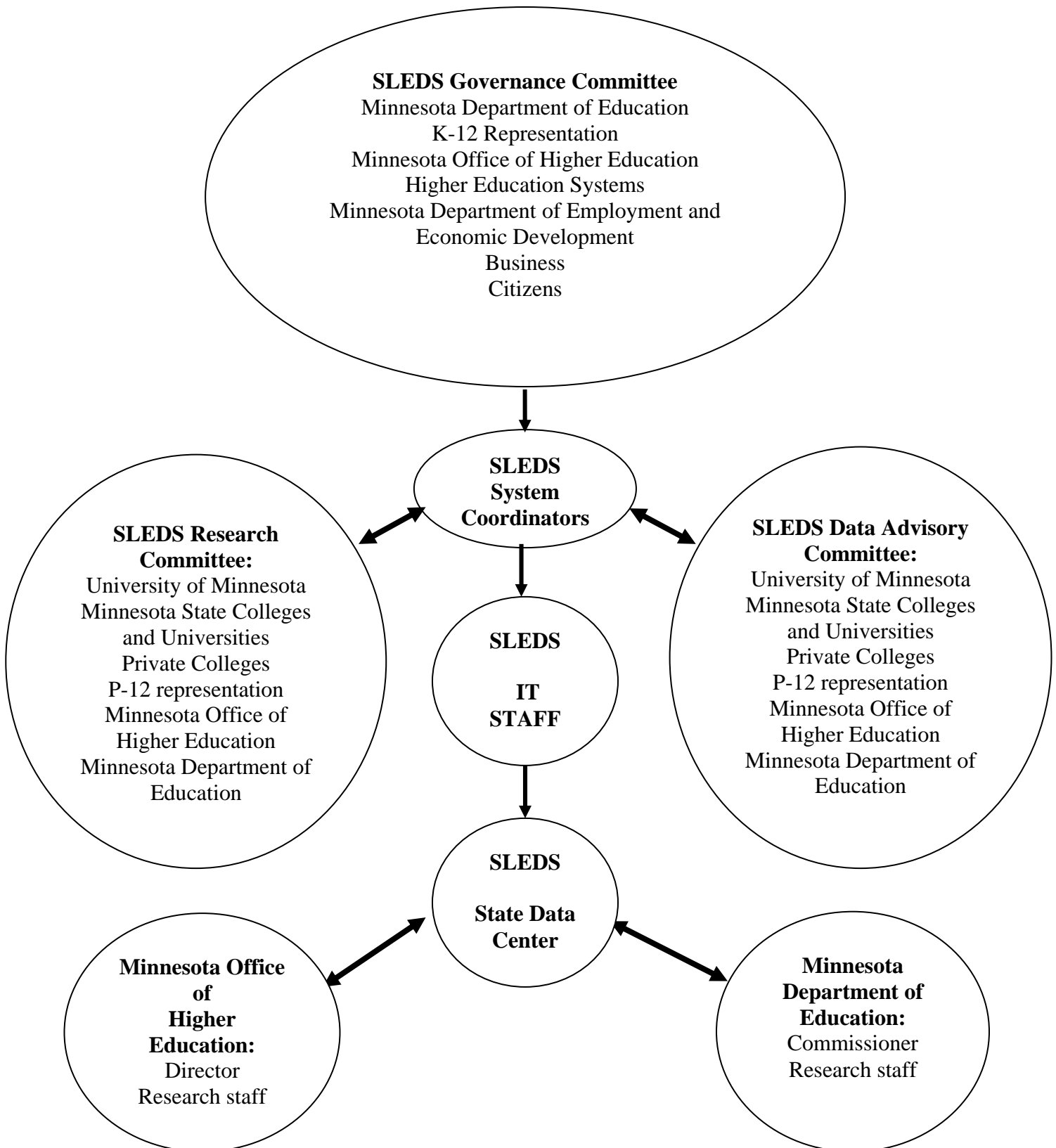
**SLEDS IT Staff** (at both Minnesota Department of Education and Minnesota Office of Higher Education) responsibilities include:

1. Assure data security protocols.
2. Construct and test required data files.
3. Manage data (set data standards, define data elements, document data processes, and identify file specifications).
4. Serve on the SLEDS Data Advisory Committee.
5. Assist in utilizing the SLEDS data to further agency research goals.
6. Manage SLEDS data system including development of end-user interfaces and automated report structures.
7. Provide technical expertise and consultation on data file construction, data linkages, and research methodologies.

**State Data Center** responsibilities include:

1. House the SLEDS data system and related server equipment.
2. Responsible for data system related issues including hardware, security, and user access.

# Data Governance Structure Relationships

**SLEDS Governance Committee**
Minnesota Department of Education
K-12 Representation
Minnesota Office of Higher Education
Higher Education Systems
Minnesota Department of Employment and
Economic Development
Business
Citizens

**SLEDS
System
Coordinators**

**SLEDS Research
Committee:**
University of Minnesota
Minnesota State Colleges
and Universities
Private Colleges
P-12 representation
Minnesota Office of
Higher Education
Minnesota Department of
Education

**SLEDS
IT
STAFF**

**SLEDS Data Advisory
Committee:**
University of Minnesota
Minnesota State Colleges
and Universities
Private Colleges
P-12 representation
Minnesota Office of
Higher Education
Minnesota Department of
Education

**SLEDS
State Data
Center**

**Minnesota Office
of
Higher
Education:**
Director
Research staff

**Minnesota
Department of
Education:**
Commissioner
Research staff

# H. Data Elements

**Baseline Student Data Variable List for the P-20 Statewide Longitudinal Education Data System (SLEDS)**

The Office of Higher Education currently has a student record data base on all students enrolled during the fall at postsecondary institutions eligible to participate in Minnesota-funded student financial aid programs. The Minnesota Department of Education has a data base on students enrolled in public schools. SLEDS would contain data from OHE and MDE for research. Below are baseline data variables each agency currently collects that will be used to populate the SLEDS and a list of recommended variables to be added in the future. **Note**: some of the variables listed below would only be used to match student records between OHE and MDE and would not be contained in the SLEDS system. The proposed structure of the SLEDS will contain a random anonymous identifier.

| Student Data Variables Currently Collected to be Used in the SLEDS | |
|---|---|
| **Minnesota Office of Higher Education** | **Minnesota Department of Education** |

**Attendance Data**

| Minnesota Office of Higher Education | Minnesota Department of Education |
|---|---|
| College attending (name) | School of attendance |
| Type of college (public, private, etc.) | School district (where student goes to school) |
| | School type (Title I eligible, charter, alternative learning center) |
| Fiscal year of data collection | Academic year |
| Term (fall enrollment only) | |
| Transfer Institution Code | |

**Student Demographic Data**

| Minnesota Office of Higher Education | Minnesota Department of Education |
|---|---|
| Name (first, middle, last) | Name (first, middle, last) |
| Birth date | Birth date |
| Gender | Gender |
| Racial/ethnic origin | Racial/ethnic origin |
| County of residence (at time of admittance) | Resident district (where student lives) |
| State of residence (at time of admittance) | Home primary language |
| Citizenship/immigration status | MARSS Student ID Number |
| MARSS Student ID Number | |

**Student Enrollment Data**

| Minnesota Office of Higher Education | Minnesota Department of Education |
|---|---|
| Student level (freshmen/sophomore/etc.) | Grade level |
| Registration type (new student/continuing/transfer, etc.) | Attendance days |
| Enrollment status (enrolled full-time or part-time) | Membership days (days a student is enrolled) |
| Degree/certificate seeking (yes/no) | Last location of attendance (determines student mobility) |
| MN high school of graduation/GED/did not graduate | Withdrawal status |
| Year of high school graduation | Graduation status |
| | Year of high school graduation |

**Academic Performance and Testing Data**

| Minnesota Office of Higher Education | Minnesota Department of Education |
|---|---|
| Credits taken | MCA test results |
| Remedial credits | ACT test scores |
| Accumulated credits earned | SAT test scores |
| Transfer credits earned | |

**Baseline Student Data Variable List for the P-20 Statewide Longitudinal Education Data System (SLEDS)**

The Office of Higher Education currently has a student record data base on all students enrolled during the fall at postsecondary institutions eligible to participate in Minnesota-funded student financial aid programs. The Minnesota Department of Education has a data base on students enrolled in public schools. SLEDS would contain data from OHE and MDE for research. Below are baseline data variables each agency currently collects that will be used to populate the SLEDS and a list of recommended variables to be added in the future. **Note**: some of the variables listed below would only be used to match student records between OHE and MDE and would not be contained in the SLEDS system. The proposed structure of the SLEDS will contain a random anonymous identifier.

| Student Data Variables Currently Collected to be Used in the SLEDS | |
|---|---|
| **Minnesota Office of Higher Education** | **Minnesota Department of Education** |
| **Academic Performance and Testing Data (cont')** | **Special Services** |
| Major 1 (program code of 1st major) | Gifted/talented participation |
| Major 2 (program code of 2nd major) | PSEO participation |
| Degree level (diploma/associate/bachelor/etc.) | Economic indicator (free/reduced lunch) |
| College graduation date | Limited English services |
| Academic award received | Title I student eligibility |
| Date academic award received | |

| Proposed NEW Student Data Variables to be Collected | |
|---|---|
| **Minnesota Office of Higher Education** | **Minnesota Department of Education** |
| Advanced standing credits granted for new students | High school core course curriculum |
| College GPA | GPA |
| | Participation in college access programs (TRIO, Admission Possible, etc.) |
| | Participation in college preparatory courses (listed individually) Post Secondary Enrollment Options (PSEO), Advanced Placement (AP), International Baccalaureate (IB), College in the Schools (CIS), Concurrent Enrollment, College Level Examination (CLEP) |
| | Class Rank |

*Updated August 19, 2009 as agreed upon by OHE and MDE*

## I. Attachments

- White paper, prepared by the Minnesota Department of Education, detailing the linking methodology to be used in creating the anonymous student links to provide data for the data warehouse.

- Data sharing agreement between the Minnesota Department of Education and Minnesota Office of Higher Education. Signed on September 1, 2009.

- Enrollment data sharing agreement between Office of Higher Education and postsecondary institutions.

- Recommendation from the College-and Career-Ready Policy Institute.

- Minnesota P-20 Education Partnership materials (description, membership roster, mission statement; *to be included when available*).

# Creating Student Information from Legacy Record Systems Using Probabilistic Record Linking

*By John Paulson, Chief Information Officer*

**November 24th, 2008**

Department
*of* Education

## Contents

## Abstract

Linking student information across disparate legacy systems within the Minnesota Department of Education (MDE) was tedious, error prone, and lacked consistent repeatable results. MDE reviewed several methods of record linking in the record linking literature. After evaluation, MDE chose to use the Howard B. Newcombe record linking techniques (Newcombe, 1988), described in the "Handbook of Record Linking". This paper describes the steps and processes used by MDE to create probabilistic record matches for one enrollment system, the Minnesota Attendance Reporting Student System (MARSS), which reports student enrollment at districts and schools. Techniques and examples are described and several applications of the results are presented.

## Introduction

The Minnesota Department of Education (MDE) collects data about students, teachers and other individuals in various separate data collection systems. In addition, MDE receives information on individuals from districts, vendors, schools, and other sources. In many instances, these systems stand alone and work well as designed. There are however, many significant requirements from various federal, state and local initiatives that need to look at this information from a student, or instructional professional perspective. Providing a student centric or teacher centric view requires linking information on individuals from these various disparate systems into a consistent view. One of the most difficult challenges related to this process is choosing linking methods for individuals that are automatable, highly reliable, repeatable, and justifiable.

## Problem Statement

MDE has extensive data repositories that have information useful to students, teachers, legislators, researchers, and state and federal accountability and reporting systems. These stake holders could create positive impact to educational outcomes if given reliable access to this repository. As MDE began creating the data warehouse repositories necessary to support this access, it became clear that one of the largest problems associated with analysis was the proper linking of data between reporting systems. If linking wasn't provided as part of the warehouse repository access, it would an exercise for each person accessing the data. This would lead naturally to multiple linking

methods which create data reporting errors and inconsistencies.

### Legacy Systems Data

Individual systems that have been designed over many years (legacy systems), have been created to serve a specific purpose. The Minnesota Automated Reporting Student System was designed, implemented and refined in the 1980s to collect information about student enrollment at the district and school level. The information it collected was used almost exclusively for financial calculations, and in fact was created and used by the Program Finance department within the agency. A separate system was developed for collecting information on Carl Perkins student program participation. Still other systems collect information regarding student disciplinary incidents, migrant status, special education statuses, graduation rates, and others. In the 70s, 80s and early 90s, object oriented techniques and database design had not evolved or been incorporated to the degree necessary to manage the complexity of creating a consistent student view. In fact, complexity was usually addressed by the creation of separate systems that prevented the inadvertent interaction between one collection and another. This created the following linking challenges:

#### Different key identifiers

Two different legacy systems that collect student information may use different student attributes to identify students. For example, a student enrollment identifier may include last name while a disciplinary incident report might not. Such inconsistencies mean that there is

less "context" information with which to link records together.

### Inconsistent use of identifiers

Some legacy systems may permit 30 characters for a last name while another permits 40 and a third allows 40 characters for last, first and middle using comma separated values. Such inconsistencies create difficulties in the automation of comparisons. Some systems even permit prefix and suffix attachments in the name. The result is that even if two legacy systems contain the same identifying attribute, the inconsistent use makes it difficult to use.

### Schedule driven inconsistencies

Legacy systems designed for a particular purpose often have a "reporting schedule" associated with them. The MARSS system collects enrollment information necessary for financial processing. While there are a number of scheduled submissions through the year, they are in preparation for processing fall enrollment financial payments and end of year enrollment financial payments. Consequently, enrollment information is accurate only twice a year when it is needed by the Program Finance department. If the NCLB and AYP systems need to measure assessment participation against enrollment, they are dependent on the financial submission schedule that does not coincide with the NCLB and AYP schedule. While MDE has addressed this particular silo problem with a system enhancement, it is an example of linking issues associated with schedule inconsistencies.

### Data Reporting Errors

If there is no central definition or standard associated with student data collection, data reporting errors may be impossible to detect. Certainly between legacy systems there is no way to know if "Anderson" and "Andersen" are the same person with a common misspelling, or if the two names reference different people. Misspelling, transpositions, the use of special characters, incorrect ordering of first and last name, and scores of other possible clerical errors contribute to data reporting errors. Independent legacy collection systems have no mechanism to enforce consistency between them.

## *Changes over Time*

Most things change over time. That includes people and systems.

### People change identifiers

People change their names due to adoption, divorce, marriage, religion, or preference. Recognizing this change between disparate reporting systems and even in the same system in separate reporting cycles represents a challenge in record linking.

### Systems evolve and change

In addition to people changing, systems change. Prior to 1997, the MARSS system used Social Security Numbers (SSNs) as student identifiers. In 1997 a policy change directed that SSN not be used and become optional. A new identifier, the MARSS number was created. This MARSS number was to be unique per student, but as will be seen later, it is only unique per student per financial reporting cycle. In 2008 another policy directive required, all

student SSNs in MDE systems be removed.

Linking records for a single student even within one system can be difficult when name changes and system changes create inconsistent identifiers.

## Previous Options

### *Manual Matching*

Humans do matching of records well, and when there is a small enough amount of data it is an acceptable alternative. We are interested in many millions of records and so this alternative was not considered except where needed to validate or audit automated methods.

### *Ad hoc Automated Matching*

This is the technique that has been used for many years at MDE. It consists of using a SQL programmer to match records according to some criteria, review the results, refine the match, and continue until a "reasonable" or "expected" match is returned. Usually if you used the same SQL programmer you would get a consistent process. But the process would need to be modified in each case to account for the legacy problems described above. While this technique worked well, it depended on scarce resources, was dependent on interpretation, and was not automatic.

The central problem with this ad hoc approach, besides the manual nature, was in the qualification or quantification of the match quality. The following questions are difficult to answer:

- How well does the match work?
- Is there bias across gender or race?
- Is there bias across highly mobile populations?
- Will I get the same answer next year?
- Where and how are the techniques used in the frequent requests documented?

## Alternatives Considered

### *Commercial Products*

One alternative is to just buy a commercial off the shelf (COTS) product to do the matching. While there are many such products on the market, most require extensive customization. Customization is necessary because COTS products have no "context" of your application. There is no inbuilt knowledge of the structure of your data or the type of data. COTS products do well with generic data, but miss the advantage of knowing that "students are related to districts", unless they have been customized for education. MDE may use COTS products in the future, but believed that understanding the "process" of linking will make the use of COTS products much more effective.

### *Other techniques*

Several other techniques were investigated including research from sources other than Newcombe (Newcombe, 1988). These included research articles, presentations and journal publications (Fellegi, 1964), (Thoburn, 2007) and (Winkler, 1993). Most of these works are powerful foundations and or derivatives of the

Newcombe work. The Necombe "Handbook of Record Linkage" was the simplest and most complete practitioner's manual available. It was a straightforward "how to" manual that allowed for novice understanding and quick analysis.

## *Context*

If the three rules of real-estate are location, location, location, the three rules of probabilistic record linking are context, context, context. While probabilistic record linking can be a complex process, using context associated with the data, significant results can be achieved while avoiding much of complexity. For example, if two files A and B each have 1,000,000 records and it is desired to find the matched records and no other information is known, that is a difficult task best left to a general tool. However, it is not the same if the files contain 1,000,000 records from students in Minnesota and the enrollment district for each student is known. Such a context would allow a much simpler and more accurate matching process. Context is one of the central reasons MDE chose to implement the matching process without COTS tools.

## Probabilistic Record Linking Overview

### *General Concepts*

General concepts needed for any discussion on matching include the following definitions in the context of this paper:

### True Linked Records

These are records that have been linked together and are verified to be correct links. They are linked and they should be linked.

### True Non-Linked Records

These are records that have not been linked together and are verified to be correct non-links. They are not linked and they should not be linked.

### False Positives

These are records that have been linked together and are found to be incorrect links. They are linked and they should not be linked.

### False Negatives

These are records that have not been linked together and are verified to be incorrect non-links. They are not linked but they should be linked.

## *Frequency Ratios*

Newcombe describes Frequency Ratios (FR) as "betting odds" in favor of a correct match. The "$E=MC^2$" of probabilistic record linking is stated as:

**Frequency Ratio**
**$=$ Frequency of Outcome $(\mathbf{x}, \mathbf{y})$ among linked pairs**
**$/$ Frequency of Outcome $(\mathbf{x}, \mathbf{y})$ among nonlinked pairs**

To illustrate an example for student matching, assume there are two files X and Y, that each contains 10,000 student records from different systems.

If you plan to probabilistically link student records from file "X" and file "Y", it is first required to create a file of "linked pairs" (x,y) where "x" is a record from file X and "y" is a record from file Y that have been determined to be about the same student. Further assume this

file is created to have 100 linked (x,y) records and is called "L".

Likewise it is also required to create a file of "unlinked pairs" (x,y) where "x" is a record from file X and "y" is a record from file Y that have been determined to be about different students. This file is also created to have 100 nonlinked (x,y) records and is designated "N".

"Outcome" is any comparison you might think valuable to measure. For this example outcome will be "Exact Match (EM) of Last Name (LN)".

The formula can then be restated as:

$$FR = Frequency\ of\ EM\ (LNx, LNy) in\ L\ /\ Frequency\ of\ EM\ (LNx, LNy) in\ N$$

If it was observed that in "L" that last name was an exact match 96 times and that last name matched 2 times in N, then the frequency ratio of "agreement" would be 96/2 or 48. In betting terms this means an exact match of last name is 48 to 1 in favor of linking. Further, it can be extrapolated that the frequency ration of "disagreement" is 4/98 or 1/24.5. In betting terms this means that a non match of last name is 24.5 to 1 against linking.

## *Global vs. Specific*

In the above example, no consideration is given to the "value" of the data. That is an exact match of "Anderson" is equal in value to an exact match of "Toqeville". When data value is ignored, the frequency ratio is considered to be a "global" frequency ratio (GFR). When data is considered and factored into the calculation the result is considered "Specific". While there is additional power in using Specific frequency ratios, it also introduces significant complexity. Not only are the formulas more complex, but the process become sensitive to the data sources. Name occurrences and frequencies will change depending on the part of the country and the name type of data being analyzed. A file of "Migrant students" will contain a different concordance listing than a locally produced file of students in a small geographic range.

MDE avoided use of specific value discrimination in our process in favor of simplification. It is mentioned here to note that if insufficient match quality is achieved with simple GFRs, they may be extended to SFRs to improve the match.

## *Conditional Probabilities*

Conditional probabilities add additional complexities. A simple example of conditional probability is matching on Last Name (LN) and Last Initial (LI). The GRFs associated with these two outcomes are not independent. If these two comparisons are needed, this may be handled two ways.

Compare conditional

LI agrees
   LN agrees compared only if LI agrees
   LN disagrees compared only if LI agrees
LI disagrees

Compare concatenated

Both LI and LN agree
LI agrees but LN does not agree
LI disagrees

Concatenated GRFs are easier to work with in following stages of calculation.

MDE avoided using conditional probabilities as much as possible and used the simpler concatenated method when necessary.

### *Partial Agreement*

Converting GFRs to SFRs when there is full agreement in an outcome is an extra step, but is not difficult. Converting GFRs to SFR when there is partial agreement becomes more difficult and is less intuitive. Since MDE has not used SFRs to date, it will not be elaborated here. Newcombe does describe the process if it is needed for increased discrimination. (Newcombe, 1988)

### *Missing Identifiers*

In general missing identifiers do not argue in favor of linkage or nonlinkage. They are neutral. There are special circumstances when this is not true however and that may include middle initial. After significant experimentation, MDE did use middle initial and treated the absence of middle initial as neutral. This assumption requires further investigation.

### *Relative Odds*

Relative Odds present an overall ranking or ordering of the quality of the match. They are achieved by multiplying individual outcomes to achieve a total sum. Returning to the previous example, let's add an exact match of first name and have two frequency ratios created.

$$FR = Frequency\ of\ EM\ (LNx, LNy) in\ L / Frequency\ of\ EM\ (LNx, LNy) in\ N$$

$$FR = Frequency\ of\ EM\ (FNx, FNy) in\ L / Frequency\ of\ EM\ (FNx, FNy) in\ N$$

Recall that the result for the first formula was 96/2 or 48. Let us suppose the result of the second formula is 96/6 or 16

This can be summarized in the following table:

| ID | Outcome | Percent | | GFR L/N |
|----|---------|---------|-----|-----|
| | | Link | Non linked | |
| LN | Agree | 96 | 2 | 048.0 |
| | Disagree | 4 | 98 | 0.040 |
| FN | Agree | 96 | 6 | 0016 |
| | Disagree | 4 | 96 | 0.042 |

Now the global frequency ratios can be multiplied to get accumulated odds

| Combined event | Calculation | Relative odds |
|----|-----|-----|
| LN and FN agree | 48*16 | 768 |
| LN agrees and FN disagrees | 48*.042 | 0.96 |
| LN disagrees and FN agrees | .040*16 | .64 |
| LN and FN disagree | .040*.042 | .002 |

With larger samples and more discriminators, the numbers get large and small quickly and become cumbersome to work with. Often people convert them to base 2 logarithms to make them easier to visualize. In addition, calculating to the three digits of precision in this example with only 100 records would not be advised. In actual practice the file of L and N would be much larger and have 1,000 or even 10,000 records and allow much greater precision.

These combined orders can be used to do matching in their own right. Often products and systems stop at this point and produce matches at some

"threshold" that is set by empirical observation.

## Absolute Odds

Newcombe describes two factors that are needed to move from relative odds to absolute odds. They are (1) the probability that a search records is indeed represented in the file being searched, and (2) the size of that file. The less likely the record is in the searched file and the larger the file, the more discrimination power is needed. These two requirements can be expresses as the following formula.

$$\textit{Absolute Odds}$$
$$= \textit{Relative Odds}$$
$$\times (\textit{number of linked search records}$$
$$\div \textit{total number of search records})$$
$$\times (1$$
$$\div \textit{total number of records being searched})$$

For our example if we assume that we are searching from file X for a match in file Y, and we have determined that 9,000 of the 10,000 students in file X are in fact in file Y then our adjustment would be $9,000/10,000 * 1/10,000 = .00009$.

Thus in our example $768*.00009 = .069$. Therefore in our example, if both last name and first name match, there is a 7% probability that it is a "correct" match. Another way to say the same thing is the match would be incorrect 93% of the time. Clearly more discrimination will be needed even when dealing with 10,000 records.

## Benefit of a cookbook approach

The Newcombe method represents a cookbook approach that is powerful yet reasonable in complexity.

## Benefit of generalized matching

The method generalizes to matching across vastly different systems if discriminatory power can be found. Discriminatory power can be quantified exactly. For example, "last reported school district" in one system can be related to "address of guardian" in a different system if needed. Additional discriminatory power can be created within any related fields.

## Benefit of context simplification

Because of context, the process of using Newcombe's method can be customized and simplified for just the ease of use and discriminatory power needed.

## Benefit of communicating absolute probabilities

The use of absolute probabilities clearly communicates risk and quality issues to end users of the matching results.

## Benefit of high quality across disparate systems

The disparate systems can be linked with high quality and certainty to provide a student view of data.

**Implementation**

*Create a reference file*

Probabilities calculated depend on the source file and target file to match. It would be possible to create a separate probability matching calculation for each system. This would likely lead to confusion and inconsistency as well as more work. The following diagram illustrates the number of calculated matching algorithms needed.

# Why a Reference Table
### Fewer, more consistent matching algorithms
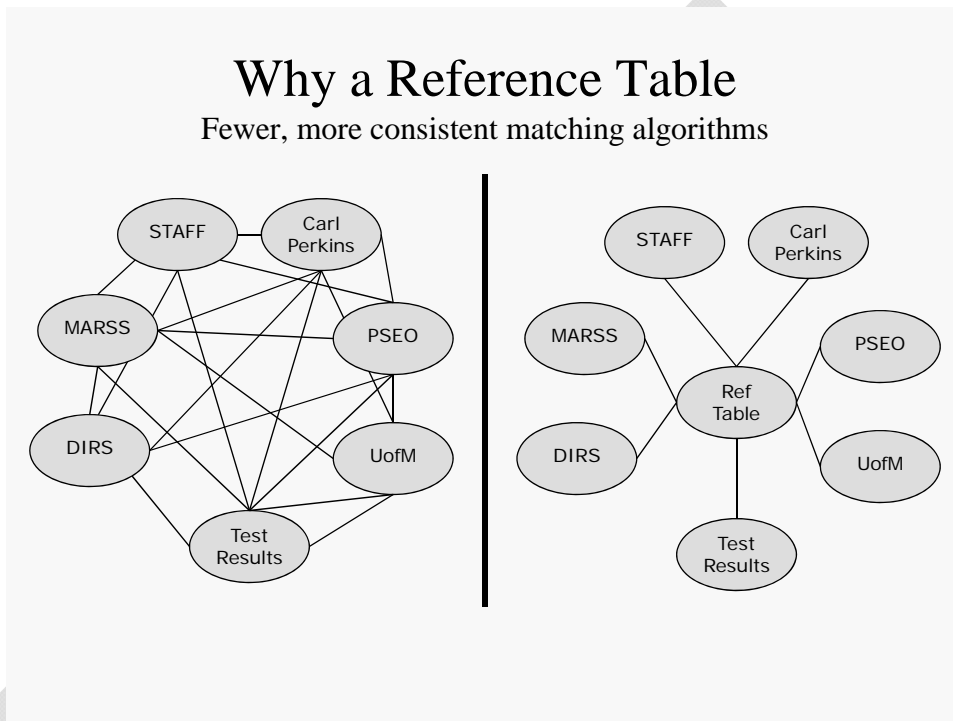


The reference file can be thought of as a file of all possible students and their aliases. If an attempt to match a record to the reference file fails, the matching process has the option of reporting the error, or adding the new record that represents the new student. The authority to add records that do not match implies the matching process that is calling for the match is "authoritative".

*Starting with Authority*

The reference table should be created with the most authoritative system available. In the case of MDE, that was the MARSS enrollment system. Districts are funded based on the data submitted and quality is monitored and audited. Since the major key field had changed from SSN to MARSS number in 1997, and our initial research was before the 2007 submission cycle completion, MDE chose to create the reference file from MARSS enrollment records from 1997 through 2006. It was also determined that the reference table initially consist of records containing MARSS number, Last Name, First Name, Middle

Initial, Gender, and Date of Birth.  Additional discriminators may be added in the future, but these were chosen as the initial set.

## *Determine Record Meta Data*

The usefulness of the resulting linking system will be enhanced if Meta data about the records being matched is available to the systems requesting linking information.  Information such as 1) from which system was this linked record derived, 2) was the system this record was derived from an authoritative system, 3) is this record a core identity record or an alias record, etc.  Meta data allows requestors to be more discriminating in their requests for record linking (e.g. "link with current authoritative records only".

### Deal with Time Upfront

It is important to note that even that matching system may change.  That means there may be calls to the system for matches that would return different results over time.  There are several ways to resolve this ambiguity.  One would be to take periodic snap shots of the reference table and "freeze" them for historical purposes.  A second would be to use versioned records that allow asking for matches as they would be at a certain point in time.  MDE chose this latter method.  Dealing with time in databases is a complete separate subject and beyond the scope of this paper.  It is important to mention however, and an excellent reference source is (Snodgrass, 2000).

## *Grouping vs. Matching*

There is a subtle distinction between grouping and matching.  Matching attempts to link records from one system to another.  Grouping is the process of creating a reference table from a single group of records that have many repetitions and aliases representing the same student.

## *Creating a reference table*

### Definition of a reference table

It is possible to link records from two separate systems, for example link file A with file B.  When several systems are involved and linking must be permitted across years and system structural changes, it is more useful to create a standard or "reference" matching file or reference table.  This table represents an evolving list of all unique persons, their aliases, and a unique identifier that describes one and only one person.  Many systems have identifiers that are more or less unique, but due to data entry errors or omissions, are not guaranteed to be unique.  The unique identifier associated with each person in the reference table is a source system independent guaranteed unique identifier.  It may be too soon to call this a "Student" table, but it is close.

### The false negative and false positive tradeoff

The reference table is created with an initial load of data from one or more systems of interest using rules determined to create the maximum true matches, while minimizing the instances of false negatives and false positives.  There is always a trade-off associated with maximum true matches and minimum false positives.  The reference table is modified operationally by those systems designated as sources of student person information.

# The Quality False Positive & False Negative Tradeoff
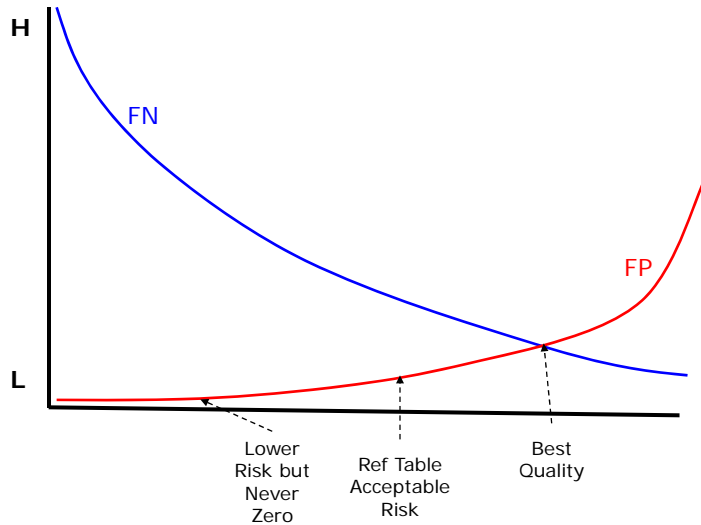


False negatives go down in number as the match is "loosened" while the opposite is true for false positives. A point where there are the fewest number of errors might be called the "Best Quality" point. MDE however took a very conservative approach. It was thought that false positives could result in the wrong information being sent to the wrong student. This was to be minimized at the sacrifice of increased false negatives. Our initial goal was to have a quality match of approximately 99.99%. This would represent one false positive per 10,000 records.

Concordance Table

| Based on 1997-2006 MARSS End of Year Data | Concordance Data |
|---|---|
| MARSS# | 1,848,400 |
| Last Name | 123,217 |
| First Name | 173,706 |
| Middle Name | 72,269 |
| Last Initial | 26 |
| First Initial | 26 |
| Middle Initial | 26 |
| Data Of Birth | 15,344 |
| Gender | 2 |

There were 11,324,620 records selected from the MDE 1997-2006 enrollment data. The concordance table above summarizes the results. Each number represents the unique number of values of that variable in the concordance table.

| Rule | Total number of records | Number of unique records after executing the rule# |
|---|---|---|
| M#, LN, FN, MN, DOB, Gender | 11324620 | 2943200 |
| M#, LN, FN, MI, DOB | 2943200 | 2570635 |
| M#, LN, FN, DOB, Gender | 2570635 | 2048318 |
| LN, FN, MN, DOB, Gender | 2048318 | 2020145 |
| LN, FN, MI, DOB, Gender | 2020145 | 2010284 |
| LN, FN, DOB, Gender | 2010284 | 1982748 |
| M#, Soundex-LN, FN, MI, DOB, gender | 1982748 | 1963508 |
| M#, LN, Soundex-FN, MI, DOB, gender | 1963508 | 1952626 |
| M#, Soundex-LN, Soundex-FN, MI, DOB, gender | 1952626 | 1952496 |
| M#, Soundex-LN, FN, MI, DOB | 1952496 | 1952433 |
| M#, LN, Soundex-FN, MI, DOB | 1952433 | 1951810 |
| M#, Soundex-LN, Soundex-FN, MI, DOB | 1951810 | 1951807 |
| M#, Soundex-LN, FN, DOB, gender | 1951807 | 1915742 |
| M#, LN, Soundex-FN, DOB, gender | 1915742 | 1909728 |
| M#, Soundex-LN, Soundex-FN, DOB, gender | 1909728 | 1907558 |
| M#, Soundex-LN, FN, DOB | 1907558 | 1907512 |
| M#, LN, Soundex-FN, DOB | 1907512 | 1907081 |
| M#, Soundex-LN, Soundex-FN, DOB | 1907081 | 1907080 |
| M#, Soundex-LN, Soundex-FN, MI, gender | 1907080 | 1888875 |
| M#, Souldex-LN, Soundex-FN not equal, MI, birthdate, gender | 1888875 | 1888464 |

(Validated by inspection of 100 aliased records for false positives.)

Creating the initial reference table load required experimentation. Retaining the conservative approach to false positives required inspection of a number of rules. The Soundex used was the Microsoft SQLServer 2005 soundex set at the highest level of four. Each rule was executed in the order shown. This allowed stronger rules to operate on larger sets of data, while weaker rules operated on smaller sets.

The initial set of records went from 11,324,620 records to 1,888,464 records with Aliases. Each group of aliases was assigned a unique id called an Alias Group ID (AGI). It is important to note that there are likely errors in this reference file, and in our case many more false negatives than false positives. Initial investigation suggests errors on the order of 100 false positives per million and 1000 false negative per million. More investigation is required to determine the exact level of error. Necombe suggests that this is not serious and will not affect the probabilities and calculations in a significant manner.

*Create a file of Linked Records*

The Newcombe method requires a file of record pairs that are known to match. MDE created a file of 10,000 records that were verified to be valid matches. The 10,000 records were selected at random from the original 11,324,620 and matched to the reference table. Each record in the file contains the "key" fields identified in the first step.

*Create a file of Non-Linked Records*

The Newcombe method requires a file of record pairs that are known to *not* match. For example, a file containing 10,000 pairs of records that have been inspected to insure that each pair of records in fact refers to different individuals (non-linked). Each record in the file contains the "key" fields identified in the first step.

*Analyze the records for Global Frequency Ratios*

The next step is to analyze the linked and non-linked files to determine Global Frequency Ratios. For example, in our "agreement" analysis, the number of times the Last Name matches on linked and non lined pairs of records can be expressed as 9385/8. This means that in 10,000 pairs of linked records the last name on both records was the same (agreed) 9385 times in the file of linked records, while the last name was the same (agreed) 8 times in the non linked record files. In looking at the "disagreement" analysis, we find 615/9992. This means that in 10,000 pairs of linked records the last name disagreed 615 times, while the last name disagreed 9992 times in the file on non lined records. This can be represented as Global Frequency Ratios (GFR) as in the following example table:

| | Comparison Outcomes | Percentage Frequencies | | Global Frequency Ratios |
| --- | --- | --- | --- | --- |
| | | Links | Non-Links | |
| MARSS# | Agree | 9289 | 0 | 9289 |
| | Disagree | 711 | 10000 | 0.0711 |
| Last Name | Agree | 9385 | 8 | 1173.125 |
| | Disagree | 615 | 9992 | 0.0615492 |
| First Name | Agree | 8665 | 12 | 722.08333 |
| | Disagree | 1335 | 9988 | 0.1336604 |
| Middle Name | Agree | 423 | 30 | 14.1 |
| | Disagree | 5862 | 6255 | 0.9371703 |
| L Initial | Agree | 9798 | 560 | 17.496429 |
| | Disagree | 202 | 9440 | 0.0213983 |

| | | | | |
|---|---|---|---|---|
| **F Initial** | **Agree** | 9960 | 694 | 14.351585 |
| | **Disagree** | 40 | 9306 | 0.0042983 |
| **M Initial** | **Agree** | 3528 | 596 | 5.9194631 |
| | **Disagree** | 6471 | 9403 | 0.6881846 |
| **Day of DOB** | **Agree** | 9814 | 345 | 28.446377 |
| | **Disagree** | 186 | 9655 | 0.0192646 |
| **Month of DOB** | **Agree** | 9913 | 889 | 11.150731 |
| | **Disagree** | 87 | 9111 | 0.0095489 |
| **Year of DOB** | **Agree** | 9906 | 524 | 18.90458 |
| | **Disagree** | 94 | 9476 | 0.0099198 |
| **Gender** | **Agree** | 9928 | 4942 | 2.0089033 |
| | **Disagree** | 72 | 5058 | 0.0142349 |
| **Sound-X of Last Name = 4** | **Agree** | 9714 | 24 | 404.75 |
| | **Disagree** | 286 | 9976 | 0.0286688 |
| **Sound-X of First Name = 4** | **Agree** | 9727 | 56 | 173.69643 |
| | **Disagree** | 273 | 9944 | 0.0274537 |
| **Sound-X of Middle Name = 4** | **Agree** | 1263 | 73 | 17.30137 |
| | **Disagree** | 4983 | 6190 | 0.8050081 |
| **Sound-X of Last Name >= 3** | **Agree** | 9744 | 367 | 26.550409 |
| | **Disagree** | 256 | 9633 | 0.0265753 |
| **Last Name not equal and Sound-X of Last Name equal** | **Agree** | 329 | 16 | 20.5625 |
| | **Disagree** | 9671 | 9984 | 0.9686498 |
| **First Name not equal and Sound-X of First Name equal** | **Agree** | 1062 | 44 | 24.136364 |
| | **Disagree** | 8938 | 9976 | 0.8959503 |
| **Middle Name not equal and Sound-X of Middle Name equal** | **Agree** | 840 | 43 | 19.534884 |
| | **Disagree** | 5406 | 6220 | 0.8691318 |
| **Last Name not equal and Sound-X of Last Name not equal and Last Initial Equal** | **Agree** | 56 | 347 | 0.1613833 |
| | **Disagree** | 9944 | 9653 | 1.0301461 |
| **First Name not equal and Sound-** | **Agree** | 75 | 363 | 0.2066116 |

| | | | | | |
|---|---|---|---|---|---|
| **X of First Name not equal and first initial Equal** | **Disagree** | 9925 | 9637 | 1.0298848 | |
| **Middle Name not equal and Sound-X of Middle Name not equal and Middle Initial Equal** | **Agree** | 927 | 171 | 5.4210526 | |
| | **Disagree** | 5358 | 6114 | 0.8763494 | |

Agreement argues for linking, disagreement argues against linking.

*Compute relative odds (weight or ranking)*

Once the Global Frequency Ratios have been recorded, relative odds can be computed.  Relative odds for any given match situation represent the product of the individual global frequency ratios.  For example, consider the following table:

| MARSS# | Last Name | First Name | Middle Name | Date of Birth | Gender | Rule Relative Probability |
|---|---|---|---|---|---|---|
| **ExactMatch** | **ExactMatch** | **ExactMatch** | **Soundex** | **ExactMatch** | **ExactMatch** | 1.85169E+15 |
| **ExactMatch** | **ExactMatch** | **ExactMatch** | **ExactMatch** | **ExactMatch** | **ExactMatch** | 1.33652E+15 |
| **ExactMatch** | **Soundex** | **Soundex** | **Soundex** | **ExactMatch** | **ExactMatch** | 1.08489E+12 |
| **ExactMatch** | **Soundex** | **Soundex** | **Soundex** | **ExactMatch** | **ExactMatch** | 1.08489E+12 |
| **ExactMatch** | **Soundex** | **Soundex** | **Soundex** | **ExactMatch** | Unknown | 5.40039E+11 |
| **ExactMatch** | **Soundex** | **Soundex** | **Soundex** | **ExactMatch** | Unknown | 5.40039E+11 |
| **ExactMatch** | **ExactMatch** | **Soundex** | **Soundex** | MD-Only | **ExactMatch** | 32477908679 |
| **ExactMatch** | **ExactMatch** | **Soundex** | **Soundex** | YD-Only | **ExactMatch** | 53003207550 |
| **ExactMatch** | **ExactMatch** | **Soundex** | **Soundex** | MD-Only | **ExactMatch** | 32477908679 |
| **ExactMatch** | **ExactMatch** | **ExactMatch** | **Soundex** | MD-Only | **ExactMatch** | 9.71636E+11 |
| **ExactMatch** | **ExactMatch** | **ExactMatch** | **Soundex** | YM-Only | Unknown | 6.24227E+11 |
| **ExactMatch** | **ExactMatch** | **ExactMatch** | **Soundex** | YD-Only | Unknown | 7.8933E+11 |
| **ExactMatch** | **ExactMatch** | **ExactMatch** | **Soundex** | MD-Only | Unknown | 4.83665E+11 |
| | | | | | | |
| **ExactMatch** | Unknown | **Soundex** | Initial | **ExactMatch** | **ExactMatch** | 15987460619 |
| Unknown | **ExactMatch** | **ExactMatch** | Unknown | **ExactMatch** | **ExactMatch** | 10204411860 |
| | | | | | | |
| **Disagree** | **ExactMatch** | **ExactMatch** | **Soundex** | **ExactMatch** | **ExactMatch** | 14173216138 |
| **Disagree** | **Soundex** | **Soundex** | **Soundex** | **ExactMatch** | **ExactMatch** | 8303946.732 |
| **Disagree** | **Soundex** | **Soundex** | **Soundex** | **ExactMatch** | Unknown | 4133572.195 |
| **Disagree** | **Soundex** | **Soundex** | **Soundex** | YM-Only | **ExactMatch** | 5623.649722 |
| **ExactMatch** | **Soundex** | **Soundex** | **Soundex** | Y-Only | Unknown | 161268.8886 |
| **ExactMatch** | **Soundex** | **Soundex** | **Soundex** | D-Only | **ExactMatch** | 487494.3004 |
| **Disagree** | **Soundex** | **Soundex** | **Soundex** | MD-Only | **Disagree** | 30.87556855 |
| **Disagree** | **Soundex** | **Soundex** | **Soundex** | **ExactMatch** | **ExactMatch** | 8303946.732 |
| **Disagree** | Initial | Initial | Unknown | **ExactMatch** | **ExactMatch** | 215067.7268 |
| | | | | | | |
| **Disagree** | **Soundex** | **Soundex** | **Soundex** | MD-Only | **Disagree** | 30.87556855 |
| **ExactMatch** | **Soundex** | **Disagree** | **Soundex** | Y-Only | **Disagree** | 12.71262266 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Disagree** | **Soundex** | Initial | **Disagree** | YM-Only | **ExactMatch** | 160.4183184 |
| **Disagree** | **Soundex** | **Disagree** | Unknown | **ExactMatch** | **Disagree** | 16.68012343 |
| **Disagree** | Initial | Initial | Initial | Unknown | **ExactMatch** | 212.3049993 |
| **Disagree** | **Disagree** | Initial | Initial | **ExactMatch** | **ExactMatch** | 4478.48211 |
| **ExactMatch** | **Disagree** | **Disagree** | **Disagree** | **Disagree** | **Disagree** | 1.8603E-06 |
| **Disagree** | **Disagree** | **Disagree** | Initial | **ExactMatch** | **ExactMatch** | 41.70937746 |

Relative Odds are useful in comparing one type of match with another to rank order probabilities. Because the numbers get quite large when dealing with more than three discriminators and quickly move to scientific notation, relative odds are often expressed as Log(2) and called weights. MDE converted directly to absolute probabilities in the next step.

*Compute absolute odds*

Relative odds are useful in comparing matching techniques. They require an arbitrary "cut-off" set by experience. Relative odds, or Global Frequency Ratios, for example can be expressed as a percentage based on the concept of number of good events divided by the number of bad events. They serve only as a rank ordering of the types of matches. Converting to absolute odds requires an additional computation. The odds of matching by "chance" need to be computed from two additional pieces of information.

$$Absolute\ Odds = Relative\ Odds * \frac{1}{TotalRecordsToSearch} * \frac{TotalLinkedRecords}{TotalSearchRecords}$$

In the MARSS reference table there are a total of 2,942,200 unique records to search. When Minnesota districts submit MARSS enrollment data, approximately 800,000 students are expected with 50,000 new students. That means that the total records we expect to link are 750,000, making our ratio 750,000/800,000 or 75/80 or 15/16. The following table shows the relative probability and the absolute probability of several rules. It may be important to note that there are many thousands of possible combinations of rules. These are some representative examples. The absolute probability is a "confidence" probability of the match.

| | MARSS# | Last Name | First Name | Middle Name | Date of Birth | Gender | Rule Relative Probability | MARSS Submission Absolute |
|---|---|---|---|---|---|---|---|---|
| MATCH | ExactMatch | ExactMatch | ExactMatch | Soundex | ExactMatch | ExactMatch | 1.85169E+15 | 99.999999830% |
| MATCH | ExactMatch | ExactMatch | ExactMatch | ExactMatch | ExactMatch | ExactMatch | 1.33652E+15 | 99.999999765% |
| MATCH | ExactMatch | Soundex | Soundex | Soundex | ExactMatch | ExactMatch | 1.08489E+12 | 99.999710623% |
| MATCH | ExactMatch | Soundex | Soundex | Soundex | ExactMatch | ExactMatch | 1.08489E+12 | 99.999710623% |
| MATCH | ExactMatch | Soundex | Soundex | Soundex | ExactMatch | Unknown | 5.40039E+11 | 99.999418672% |
| MATCH | ExactMatch | Soundex | Soundex | Soundex | ExactMatch | Unknown | 5.40039E+11 | 99.999418672% |
| MATCH | ExactMatch | ExactMatch | Soundex | Soundex | MD-Only | ExactMatch | 32477908679 | 99.990334630% |
| MATCH | ExactMatch | ExactMatch | Soundex | Soundex | YD-Only | ExactMatch | 53003207550 | 99.994077288% |
| MATCH | ExactMatch | ExactMatch | Soundex | Soundex | MD-Only | ExactMatch | 32477908679 | 99.990334630% |
| MATCH | ExactMatch | ExactMatch | ExactMatch | Soundex | MD-Only | ExactMatch | 9.71636E+11 | 99.999676895% |
| MATCH | ExactMatch | ExactMatch | ExactMatch | Soundex | YM-Only | Unknown | 6.24227E+11 | 99.999497074% |
| MATCH | ExactMatch | ExactMatch | ExactMatch | Soundex | YD-Only | Unknown | 7.8933E+11 | 99.999602270% |
| MATCH | ExactMatch | ExactMatch | ExactMatch | Soundex | MD-Only | Unknown | 4.83665E+11 | 99.999350916% |
| | | | | | | | | |
| REVIEW | ExactMatch | Unknown | Soundex | Initial | ExactMatch | ExactMatch | 15987460619 | 99.980367132% |
| REVIEW | Unknown | ExactMatch | ExactMatch | Unknown | ExactMatch | ExactMatch | 10204411860 | 99.969244207% |
| | | | | | | | | |
| REVIEW | Disagree | ExactMatch | ExactMatch | Soundex | ExactMatch | ExactMatch | 14173216138 | 99.977854581% |
| REVIEW | Disagree | Soundex | Soundex | Soundex | ExactMatch | ExactMatch | 8303946.732 | 72.565633561% |
| REVIEW | Disagree | Soundex | Soundex | Soundex | ExactMatch | Unknown | 4133572.195 | 56.834599476% |
| REVIEW | Disagree | Soundex | Soundex | Soundex | YM-Only | ExactMatch | 5623.649722 | 0.178810289% |
| REVIEW | ExactMatch | Soundex | Soundex | Soundex | Y-Only | Unknown | 161268.8886 | 4.885925933% |
| REVIEW | ExactMatch | Soundex | Soundex | Soundex | D-Only | ExactMatch | 487494.3004 | 13.441045365% |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| INSERT | Disagree | Soundex | Soundex | Soundex | MD-Only | Disagree | 30.87556855 | 0.000983472% |
| REVIEW | Disagree | Soundex | Soundex | Soundex | ExactMatch | ExactMatch | 8303946.732 | 72.565633561% |
| REVIEW | Disagree | Initial | Initial | Unknown | ExactMatch | ExactMatch | 215067.7268 | 6.411356120% |
| | | | | | | | | |
| INSERT | Disagree | Soundex | Soundex | Soundex | MD-Only | Disagree | 30.87556855 | 0.000983472% |
| ERROR | ExactMatch | Soundex | Disagree | Soundex | Y-Only | Disagree | 12.71262266 | 0.000404935% |
| INSERT | Disagree | Soundex | Initial | Disagree | YM-Only | ExactMatch | 160.4183184 | 0.005109557% |
| INSERT | Disagree | Soundex | Disagree | Unknown | ExactMatch | Disagree | 16.68012343 | 0.000531311% |
| INSERT | Disagree | Initial | Initial | Initial | Unknown | ExactMatch | 212.3049993 | 0.006762112% |
| REVIEW | Disagree | Disagree | Initial | Initial | ExactMatch | ExactMatch | 4478.48211 | 0.142450261% |
| ERROR | ExactMatch | Disagree | Disagree | Disagree | Disagree | Disagree | 1.8603E-06 | 0.000000000% |
| INSERT | Disagree | Disagree | Disagree | Initial | ExactMatch | ExactMatch | 41.70937746 | 0.001328555% |

MDE uses the probabilities to decide what to do with a record. If the record matches on a very high probability rule, say 99.99% or better, then that is called a match and the student is known. If the record does not match on any probability below 0.00, then that record is considered a new student and can be added to the reference table if the matching system is authoritative. If the record submitted matches at some rate in-between, it is considered a grey match and must be reviewed.

Of course this scale can be changed per process and each matching program is free to set their unique confidence requirements.

Also note that there are a couple of error conditions noted. This occurs when the student number is found in the reference table, but the data associated with that number does not match. The wrong number has been associated with the student data submitted.

*Choosing Rules*

It is theoretically possible to attempt to match each record starting with the best rule and working downward until the highest possible match is obtained. In practice, since there are many thousands of possibilities and millions of records, the performance associated with the dynamic calculations is prohibitive. MDE chose to experiment with matching rules and use a subset for processing.

Default Matching Rules

These are the suspect matching rules used by the SLS API as of 1-23-08.

| MARSS# | Last Name | First Name | Middle Name | Date of Birth | Gender | MARSS Submission Absolute |
|---|---|---|---|---|---|---|
| **ExactMatch** | **Soundex** | **Soundex** | Unknown | **ExactMatch** | Unknown | 99.9886% |
| Unknown | **ExactMatch** | **ExactMatch** | Unknown | **ExactMatch** | **ExactMatch** | 99.9692% |
| **ExactMatch** | **ExactMatch** | **ExactMatch** | Unknown | Unknown | **ExactMatch** | 99.9801% |
| **ExactMatch** | Unknown | **Soundex** | Initial | **ExactMatch** | **ExactMatch** | 99.9804% |
| **ExactMatch** | **Soundex** | Unknown | Initial | **ExactMatch** | **ExactMatch** | 99.9770% |

Suspect Matching Rules

These are the suspect matching rules used by the SLS API as of 1-23-08. This is after the SLS Default Match Rules are executed and none of them rules create a match.

| MARSS# | Last Name | First Name | Middle Name | Date of Birth | Gender | MARSS Submission Absolute |
|---|---|---|---|---|---|---|
| **ExactMatch** | Unknown | Unknown | Unknown | Unknown | Unknown | 0.295% |
| Unknown | **Soundex** | **Soundex** | Unknown | **ExactMatch** | Unknown | 48.665% |
| Unknown | **Soundex** | **Soundex** | Unknown | YM-Only | **ExactMatch** | 0.129% |
| Unknown | **Soundex** | **Soundex** | Unknown | YD-Only | **ExactMatch** | 0.163% |
| Unknown | **Soundex** | **Soundex** | Unknown | MD-Only | **ExactMatch** | 0.100% |

## Example Successes

*Test Editing*

Districts are allowed to edit some limited information regarding test results before the information is summarized and sent to parents. Because of the significant requirement not to send the wrong test to a

parent, the test editing system used very rigid match criteria to associate students with assessment results. Districts were allowed to see the matches and correct student identifying data when matches could not be made. No automatic matches were made with confidence less than 99.99%

### *AYP Participation*

AYP used the matching program to associate enrollment records to test records. This increased visibility and accuracy of the AYP participation calculation.

### *More inclusive of highly mobile groups*

The new matching algorithms removed significant bias against the highly mobile populations. Greater accuracy of matches across LEP, FRP, race and ethnicity are now possible. Highly mobile groups were the groups most likely to be assigned multiple MARSS numbers and have spelling changes in their names between district systems. The PRL matching algorithms create better longitudinal matches.

### *Increased cohort cohesion*

The ability to have cohorts span ten or more years with highly reliable matching will permit increased insight into educational processes. An example is the exit code study.

### *Growth score calculations*

Being able to look longitudinally backwards to find prior year scores is essential for growth modeling. With PRL it is possible to look backward into prior year data without bias against highly mobile populations.

## Coming Soon

### *Ability to share with higher Ed*

Using the PRL methodology and specific contexts, MDE plans to match student information across Minnesota state agencies for the purpose of longitudinal studies. Sharing with higher education institutions will allow college preparedness studies to analyze the effectiveness of various programs and course taking patterns.

### *Ability to share with wage information*

Wage information can be used as one measure of outcome success. Crossing student information with State wage and income information will allow longitudinal studies to focus on how well Minnesota education prepares students for the work force.

### *Ability to share with human services information*

Identifying students who should have access to a beneficial program based on their qualifying in some other program, may help to identify opportunities to promote Minnesota state services where needed. Matching students in schools and districts to human services programs can remove significant burden from the districts to qualify students for basic programs assistance. Faster, more accurate, and more inclusive program administration will be possible with automated student and human services matching.

## Summary

There are significant studies and research efforts remaining, but MDE is already benefiting from the increased quality and ease of use regarding matching, especially across state agencies. The method shows great promise in longitudinal studies as well as linking disparate systems. The Necombe techniques are not new. They have been applied extensively in health care for patient record linking. Extending this technique into the educational sector seems straight forward and timely.

## Bibliography

Fellegi, I. P. (1964). A Therory for Record Linkage. *Journal of the American Statistical Association* , 1183-1210.

Fernandes, L. (2008). Patient Identification in Three Acts. *Journal of AHIMA 79, no.4, April* , 46-49.

Newcombe, H. B. (1988). *Handbook of Record Linkage Methods for Health and Statistical Studies, Administration, and Business.* Oxford: Oxford University Press.

O'Connor, Michele. (2006, August 16). *The Advantage of Algorithmic Matching.* Retrieved from National Alliance for Health Information Technology: http://www.nahit.org

Snodgrass, R. T. (2000). *Developing Time-Oriented Database Applications in SQL.* Morgan Kaufmann.

Thoburn, K. K. (2007, March 30). *Fundamentals of Linking Public Health Data Sets (a presentation) Link Plus Probabilistice Record Linking Software.* Retrieved 2007, from Center for Disease Control CDC - National Association of Health Data Associations: http://nahdo.org/cs/media/p/182.aspx

Winkler, W. E. (1993). *Matching and Record Linkage.* Washington D.C.: U.S. Bureau of the Census.

## STATE OF MINNESOTA
## INTERAGENCY DATA SHARING AGREEMENT

The parties to this agreement are the **Minnesota Department of Education (MDE)** and the **Minnesota Office of Higher Education (OHE).**

1. MDE is the state agency responsible for gathering data and conducting evaluations of K-12 education programs in Minnesota. It also is the state agency responsible for enforcing and ensuring reporting requirements on various K-12 education programs, both on a statewide basis and on behalf of school districts in Minnesota. MDE has state and federal legal authority to conduct these functions.

2. OHE is the state agency responsible for collecting data on students attending higher education institutions and pursuing postsecondary education. OHE has corresponding authority and responsibility to evaluate postsecondary education in Minnesota.

3. Pursuant to Minn. Stat. § 13.32, subd. 11, MDE and OHE also have authority to share educational data in order to analyze instruction in school districts for the purposes of improvement. Minnesota Statutes § 13.32, subd. 11, authorizes the agencies to share attendance data, student demographic and enrollment data, academic performance and testing data, and special academic services received by a student.

4. Federal law further permits this agreement. MDE and OHE each are state educational authorities under 34 C.F.R. § 99.31(a)(3)(iv). MDE and OHE are authorized by state and federal laws to access education records in order to conduct evaluations of Federal or State supported education programs, as required by 34 C.F.R. § 99.35(a)(2). MDE and OHE have in place systems that satisfy the requirements of 34 C.F.R. § 99.33(b). Under federal law found at 34 C.F.R. §§ 99.35(a)(1) and 99.33(b), authorized state educational authorities may redisclose data on behalf of local education agencies. This agreement establishes the parameters for redisclosing data, on behalf of local education agencies, from MDE to OHE and from OHE to MDE for the purpose of conducting authorized evaluations, pursuant to 34 C.F.R. § 99.35(a)(1), and for the purpose of conducting research to improve instruction, as authorized by 34 C.F.R. § 99.31(a)(6)(i)(C) and 99.31(a)(6)(ii).

5. Under this agreement, shared data shall form the Minnesota Educational Longitudinal Data System (LDS). OHE and MDE will share lists of students with data specified in Attachment A. MDE and OHE will generate a random anonymous identifier for each student to be used when sharing data. The parameters of the data sharing for purposes of creating a random anonymous identifier for each student are as follows:

(a) MDE and OHE will create a random anonymous identifier for each student for whom the agencies seek to share data. MDE and OHE will share the following data in limited numbers to validate the key generation:
      (1) data key and applicable code;
      (2) random anonymous identifier created for the student;

(3) first name, middle initial (if known), and last name of the student;

(4) data of birth; and

(5) MARSS number, high school of graduation, and year of graduation.

(b) MDE and OHE will use the random anonymous identifier to match received data with education records contained in its databases. After matches have been verified, each party to the agreement will retain only the randomized identifier. OHE will destroy all other data that personally identifies a student provided by MDE. MDE will destroy all other data that personally identifies a student provided by OHE.

(c) This matching to assign a random anonymous identifier will be completed on a periodic basis, but at least yearly, or whenever updating of the student identifiers is necessary to facilitate sharing of student data for the purposes outlined in this agreement.

6. MDE and OHE will use the random anonymous identifier created pursuant to Paragraph 5 above to share data for the purpose of analyzing data to improve instruction, and as described in Attachment A to fulfill evaluation and reporting requirements, and on behalf of school districts. The data that is subject to this agreement will not be shared for any other purposes. All modifications to Attachment A must be agreed upon in writing by both parties to the agreement.

7. All data shared pursuant to this agreement will only be transmitted by a secured method that is agreed upon by both agencies.

8. MDE and OHE will retain the data shared pursuant to this agreement in a secure manner consistent with the provisions of this agreement, except to the extent that this agreement requires the parties to the agreement to destroy data shared. MDE and OHE agree to amend their record retention policies, if necessary, to allow for destruction of the matching program data after it has been used.

9. MDE and OHE will document data exchanges under this agreement. A documentation log must include the date, description of data shared, purpose for sharing the data, and the name of the party to the agreement receiving the data.

10. MDE and OHE understand that records and information maintained by either party regarding any person may include private data and shall be protected from unauthorized use and/or disclosure under this agreement.

11. MDE and OHE agree to comply with all applicable federal and state laws, statutes, and rules with respect to the protection of privacy, security and dissemination of the shared data. Nothing in this agreement may be construed to allow either party to maintain, use, disclose or share student information in a manner not allowed by federal or state laws. MDE and OHE understand that personally identifiable information maintained by either party to the agreement is subject to the privacy and confidentiality provisions of federal and state statutes, rules and regulations, including, but not limited to, the Family Education Rights and Privacy Act (20 U.S.C 1232g); related federal regulations (34

C.F.R. Part 99); the Minnesota Government Data Practices Act, Minnesota Statutes 13.01 *et seq*; and federal laws and regulations regarding students with disabilities (20 U.S.C. §1417 (c); 34 C.F.R. 300.32, 34 C.F.R. §§ 300.610-300.627)).

12. MDE and OHE certify that all persons having access to any data shared or created under this agreement will be informed of the sensitive nature of the information and will be trained in the proper data handling and safeguard procedures.

13. All employees, contractors and agents of MDE and OHE will comply with all applicable federal and state laws with respect to the data shared under this agreement. MDE and OHE further certify that all personnel, including contractors or agents, having access to data under this agreement have been instructed regarding the governing privacy and data practices provisions. Nothing in this paragraph authorizes sharing data provided under this agreement with any other entity that is not a party to this agreement.

14. All data obtained pursuant to this Agreement will be maintained in a secure environment. All copies of data of any type, including any modifications or additions to data from any source that contains information regarding individual students, are subject to the provisions of this agreement in the same manner as the original data.

15. MDE and OHE will only disclose data in summary or aggregate form for reporting purposes.

16. All data shared pursuant to this agreement will be destroyed by MDE and OHE when it is no longer needed for the purposes for which it was shared.

17. No fees will be charged or exchanged by either MDE or OHE pursuant to this agreement.

18. Neither MDE or OHE may assign its obligations under this Agreement, nor any part of its interest in this Agreement, to another agency.

19. MDE and OHE represent that they are authorized to bind to the terms of this contract, including confidentiality and destruction or return of student data, all related or associated institutions, individuals, employees or contractors who may have access to the data or may own, lease or control equipment or facilities of any kind where the data is stored, maintained or used in any way.

20. MDE and OHE agree that they are responsible for their own acts and the results thereof to the extent authorized by law and shall not be responsible for the acts of the other party to the agreement and the results thereof. The liability of a state agency is governed by the provisions of the Minnesota Torts Claims Act, Minn. Stat. § 3.732 and 3.736, *et. seq.,* and other applicable law.

21. MDE or OHE may choose to terminate this agreement if it is deemed no longer necessary due to the creation of a different data sharing mechanism, or changes in state

statute or federal law. Either MDE or OHE may terminate this agreement at any time, with or without cause, upon 30 days written notice to the other party to the agreement.

22. This agreement becomes effective upon signature of the authorized representatives for MDE and OHE, the last date of signature becoming the effective date and will remain **in effect until December 31, 2014,** or until it is superseded by a new data sharing agreement or the creation of a different data sharing mechanism, whichever occurs first.

23. All data provided by OHE pursuant to this Agreement shall confirm to data sharing provisions under the *Enrollment Data Sharing Agreement* between OHE and the applicable institutions.

24. MDE and OHE designate a single authorized representative for purposes of maintaining the data sharing agreement and ensuring that it is properly enforced.

MDE authorized representative is <u>Alice Seagren, Commissioner, 1500 Highway 36 West, Roseville, MN 55113, (651) 582-8669.</u>

OHE authorized representative is <u>David Metzen, Director, Minnesota Office of Higher Education, 1450 Energy Park Drive. Suite 350, Saint Paul, MN 55108, (651) 259-2962.</u>
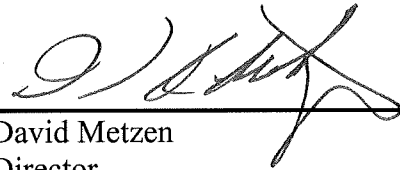
Minnesota Department of Education:　　　　Minnesota Office of Higher Education:


Alice Seagren　　　　　　　　　　　　　　David Metzen
Commissioner　　　　　　　　　　　　　　Director


9/01/09　　　　　　　　　　　　　　　　9/01/09
Date　　　　　　　　　　　　　　　　　　Date

**STATE OF MINNESOTA**
**INTERAGENCY DATA SHARING AGREEMENT**
**ATTACHMENT A**
**MINNESOTA EDUCATIONAL LONGITUDINAL DATA SYSTEM (LDS)**


The parties to this agreement are the Minnesota Department of Education (MDE) and the Minnesota Office of Higher Education (OHE).

MDE and OHE will only share data in the situations outlined in this attachment and as authorized in the signed, attached interagency data sharing agreement. Studies conducted using data from the Minnesota Educational Longitudinal Data System (LDS) under this agreement will be approved in writing by the LDS Research Committee and the LDS Governing Body.

MDE and OHE will use specified personal identifiers to assign a random anonymous identifier to each student record for the purposes of creating the Minnesota Educational Longitudinal Data System (LDS).  Categories of data elements are listed below for the purposes of research and evaluation studies to examine the transition of students from high school to postsecondary institutions. OHE and MDE will use student data to develop summary district, consortium, and state reports using only aggregate information.


**Categories of data elements to be shared as part of the LDS as allowed under Chapter 298: Sec.2. *M.S.* 2006, section 13.32 Subd.11.**

*Attendance data, including name of school or institution, school district, year or term of attendance, and term type;*

> ATTENDANCE DATA
> Data elements in this section may be used to identify and locate educational institutions which a student has attended or is attending. In the case of secondary schools, this generally refers to the school attended last or from which the student graduated. For postsecondary institutions, identification information can be included for any institution which the student previously attended, or which awarded the student a degree, diploma, or certificate; or from which transfer award units have been accepted by the institution currently attended. Generally, the term "institution" refers to the organization offering educational programs and/or instruction to students. This data also includes the year and academic term of attendance.

*Student demographic and enrollment data;*

> STUDENT DEMOGRAPHIC DATA
> Data elements in this section can be used to identify a person, (e.g., a student, his/her parents, or his/her spouse) and to describe various personal characteristics of that individual (e.g. race, gender, age).

STUDENT ENROLLMENT DATA
The data elements of this section may be used to describe the process by which a student enters an institution and/or subdivision of the institution, a process—frequently including the payment of tuition and/or fees—which results in the student's name being entered into the rolls, records, and/ or files of the institution. Data elements in this section may also be used to provide information about a student's aspirations, with respect to future education and career. Educational aspirations are expressed by, the type of formal award a student seeks or his/her objectives in attending a postsecondary institution.

*Academic performance and testing data*;

ACADEMIC PERFORMANCE AND TESTING DATA
Data elements in this section may be used to describe various aspects of a student's activities and accomplishments which are directly related to educational programs of the institution. Included are terms which describe the courses taken by the student, such as course name, grades (marks), and award units received for successful completion of courses. Also included are standardized test data.

*Special academic services received by a student*.

SPECIAL ACADEMIC SERVICES
Data elements in this section may be used to describe activities whose primary purpose is to contribute to students' emotional and physical well-being and to their intellectual, cultural, and social development outside the context of the formal instructional program. Included are elements which indicate the student's participation in gifted and talented programs, concurrent enrollment, special education, free and reduced price lunch, limited English language programs, and supplemental services.

An example of a data sharing agreement between a Minnesota postsecondary institution (University X) with the Minnesota Office of Higher Education to share student enrollment records.

# ENROLLMENT DATA SHARING AGREEMENT
## Between University X and
## the Minnesota Office of Higher Education

The University X is an educational agency or institution subject to the Family Educational Rights and Privacy Act, 20 U.S.C. 1232g, and 34 C.F.R. 99 (FERPA) and the Minnesota Government Data Practices Act, Minnesota Statutes Chapter 13 (MGDPA). This Agreement sets forth the conditions and provisions of FERPA and the MGDPA specifically applicable to University X's disclosure of personally identifiable education records (education records) to the Minnesota Office of Higher Education (OHE) in furtherance of its statutory responsibilities stated at Minn. Stat. § 136A.01, subd. 2 and for other appropriate purposes by or on behalf of University X including, but not limited to, the Minnesota Education P-20 Longitudinal Data System Project (LDS Project).

The parties acknowledge and agree that all education records provided by University X to OHE pursuant to this Agreement, and all data created, collected, received, stored, used, maintained, or disseminated in accordance with this Agreement, are subject to applicable privacy laws, including FERPA and the MGDPA.

Pursuant to this Agreement, University X shall provide education records to OHE described in the attached Appendix A and such additional education records as may be added by the LDS Governing Board to OHE as may be necessary for the following purposes: 1) in connection with financial aid, as permitted by 20 U.S.C. § l232g(b)(I)(D) and 34 C.F.R. § 99.31(a)(4); 2) subject to 20 U.S.C. § 1232g (b)(3) and 34 C.F.R. § 99.35 to an educational authority to audit or evaluate Federal or State supported education programs, or for the enforcement of or compliance with Federal legal requirements that relate to those programs; and 3) to an organization to conduct studies for or on University X's behalf, as permitted by 20 U.S.C. § 1232g(b)(I)(F) and 34 C.F.R. § 99.31(a)(6). OHE may use the education records only for the purposes for which University X provided the records or as may be subsequently authorized.

University X and the OHE specifically understand and agree that education records provided by University X to OHE under this Agreement are subject to the following terms and conditions

      1.     Any use or disclosure of University X's education records by OHE, its employees, agents or contractors is subject to and shall be consistent with applicable provisions of FERPA and the MGDPA including, but not limited to, FERPA regulations at 34 C.F.R. § 99.33 and 34 C.F.R. § 99.32 regarding recordkeeping and re-disclosure of education records.

      2.     For education records provided for the LDS Project pursuant to 34 C.F.R. § 99.31(a)(6)(i), the parties agree that the LDS Project will provide educators and policymakers with more comprehensive data and analysis from which to make informed decisions leading to educational improvement at all

levels, that it is of indefinite duration, and that the information to be disclosed is described in this agreement; and OHE shall conduct the study in a manner that does not permit personal identification of parents or students by individuals other than representatives of OHE, and the information must be destroyed when no longer needed for the purposes of the study, as required by 34 C.F.R. 99.31(a)(6)(ii).

3.  OHE shall develop, implement, maintain and use appropriate administrative, technical and physical security measures to preserve the confidentiality, integrity and availability of all electronically maintained or transmitted education records received from or on behalf of University X. These measures will be extended by contract to all subcontractors used by OHE.

4.  University X shall be a member of the LDS Governing Board that will approve in advance data security protocols, data transfer procedures and research and evaluation proposals involving any use, disclosure, re-disclosure, retention and destruction of University X education records for this project.

5.  Where appropriate, de-identification of University X's education records for the LDS Project shall be consistent with 34 C.F.R. § 99.31 (b).

6.  If the U.S. Department of Education determines that OHE has violated an applicable provision of FERPA regarding re-disclosure or requirement that the education records provided be destroyed when no longer needed for the purposes for which the records were disclosed, University X may not permit OHE access to personally identifiable information from its education records for at least five years.

7.  This Agreement may be amended in writing at any time by mutual consent of the parties. Either party may terminate this Agreement for any reason upon 90 days' written notice. Absent termination, this Agreement shall continue in effect for a period of five years.

| | |
|---|---|
| _____ | _____ |
| Date | XX, Director |
| | Minnesota Office of Higher Education |
| | |
| _____ | _____ |
| Date | XX for University X |

## Appendix A

OHE may request from University X the following education records as may be necessary for appropriate purposes:

1. Full name
2. Social security number
3. Gender
4. Birth date
5. Racial/ethnic origin
6. Place of permanent residence by state/territory and county (if Minnesota)
7. Minnesota postsecondary institution
8. Institution student identification number
9. Tuition category (resident, non-resident, reciprocity)
10. Secondary school completion status
11. Minnesota high school and year of graduation
12. Other Minnesota postsecondary institutions attended and dates and terms of attendance
13. Student type (regular or extension)
14. Student level (freshman, sophomore, etc )
15. Units of instruction and quarter-hour equivalence (regular, extension, accumulated, remedial, and transfer)
16. Disability category
17. Citizenship status
18. Postsecondary degree seeking status
19. Registration or administrative status
20. Enrollment status (full- or part-time)
21. Major(s) or program of study
22. Formal academic degree or award and date awarded
23. MARSS Student ID Number
24. Degree/certificate (yes/no)
25. Advanced standing credits granted for new students
26. Major 1 (program code of 1st major)
27. Major 2 (program code of 2nd major)

**College and Career Readiness Policy Institute recommendations for developing a P-20 Data System**: One of five subcommittees was assigned the job of developing policy recommendations for developing statewide data systems for review by a statewide leadership team and national partners.

<div align="center">

**Minnesota College and Career Policy Institute**
**Policy Recommendations from Group 5**
**October 29, 2009**

</div>

**Goal:** Minnesota will develop a longitudinal student data system to provide educators and policymakers with more comprehensive data and analysis from which to make informed decisions leading to educational improvement at all levels. As required by the federal government as a condition of receiving federal fiscal stabilization funds, this statewide longitudinal data system will match student data from pre-kindergarten through completion of postsecondary, enabling educators and policymakers to answer a range of program and policy questions to gauge the effectiveness of programs and design targeted improvement strategies.

The purpose of the longitudinal data system is to identify the predictors of long-term prekindergarten through higher education student success. The plan is to link data on students who graduate from a Minnesota public high school and attend a Minnesota postsecondary institution at the undergraduate level (approximately 40,000 high school graduates annually).

**Policy Recommendation #1  Governance:**  We recommend the P-20 Partnership  consider and adopt the P-20 Longitudinal Data System charter document and accept its role as the collaborative governing body for the system.

   a.  The P-20 Partnership will name of subset of its members as the governing board for the Longitudinal Data System including representatives of business, citizens, the Department of Employment and Economic Development, higher education systems, school districts, the Minnesota Office of Higher Education and the Minnesota Department of Education.

   b.  The governing board of the Longitudinal Data System will name members to a research committee and a data advisory committee and will clarify roles, responsibility and staffing.

   c.  The governing board of the Longitudinal Data system will approve a process for gathering statewide input from educators and policymakers on the key questions to be posed to Minnesota's LDS system.

**Policy Recommendation #2 Sustainability:**  We recommend that the P-20 Partnership amend its bylaws to accept the governance responsibilities for the Longitudinal Data System outlined in the charter and seek sufficient resources for a collaborative staffing and maintenance arrangement at the Minnesota Department of Education, the Minnesota Office of Higher Education, the Minnesota Department of Employment and Economic Development and the Office of Enterprise Technology.

**Policy Recommendation #3 Accountability:**  The Commissioner of the Minnesota Department of Education and the director of the Minnesota Office of Higher Education are jointly and

ultimately responsible for the work product, effectiveness, privacy and security of Minnesota's Longitudinal Data System.   We support consideration of policy recommendations outlined in Children's Educational Records and Privacy: A study of elementary and secondary school state reporting systems, released by Fordham Law School Center on Law and Information Policy on October 28, 2009.

**College and Career-Ready Policy Institute Members–Group 5**

| Name | Title | Organization |
| --- | --- | --- |
| Tom Baden Jr. | | Office of Enterprise Technology |
| Laura Bloomberg | Associate Director | Center for School Change |
| Chris Buse | Chief Information Security Officer | Office of Enterprise Technology |
| Emmanuel Dolo | Research Director | Minnesota Minority Education Partnership |
| Alexandra Djurovich | Policy Analyst | Minnesota Office of Higher Education |
| Meredith Fergus | Policy Analyst | Minnesota Office of Higher Education |
| Aimee Guidera | Director | Data Quality Campaign |
| Richard Howard | Director of Institutional Research | University of Minnesota |
| Nathan Lassila | Vice-President for Research & Policy Development | Minnesota Private College Council |
| Kristine Legler Kaplan | Assistant General Counsel | Minnesota State Colleges and Universities |
| John Kellogg | Associate Director, Office of Institutional Research | University of Minnesota |
| Tom LaBounty | Coordinator of Research & Testing | Rochester Public Schools |
| Mark Misukanis | Director of Fiscal Policy and Research | Minnesota Office of Higher Education |
| Pat Olson | Assistant Commissioner | Minnesota Department of Education |
| Eric Nesheim | Executive Director | Minnesota Literacy Council |
| Wendie Palazzo | Career and Technical Education Director | Minneapolis Public Schools |
| Barb Schlaefer | Director of Communications | Minnesota Office of Higher Education |
| Craig Schoenecker | System Director of Research | Minnesota State Colleges and Universities |
| Cathy Wagner | Director, Information Technology | Minnesota Department of Education |
| Vicki Weinberg | Education Specialist 2 | Minnesota Department of Education |