12-1-2010

Modeling a Longitudinal Relational Research Data System

Michelle D. Hunt Olsen

MODELING A LONGITUDINAL RELATIONAL RESEARCH DATA SYSTEM

By

Michelle D. Hunt Olsen
Bachelor of Science, Montana State University, 1979
Master of Education, Lesley College, 1994
Master of Education, Idaho State University, 1999

A Dissertation

Submitted to the Graduate Faculty

of the

University of North Dakota

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Grand Forks, North Dakota
December
2010

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

This dissertation, submitted by Michelle D. Hunt Olsen in partial fulfillment of the requirements for the Degree of Doctor of Philosophy from the University of North Dakota, has been read by the Faculty Advisory committee under whom the work has been done and is hereby approved.

Richard A. Landry
Chairperson

This dissertation meets the standards for appearance, conforms to the style and format requirements of the Graduate School of the University of North Dakota, and is hereby approved.

Joseph D. Benoit
Dean of the Graduate School

December 17, 2010
Date

PERMISSION

Title Modeling a Longitudinal Relational Research Data System.

Department Teaching and Learning

Degree Doctor of Philosophy

In presenting this dissertation in partial fulfillment of the requirements for a graduate degree from the University of North Dakota, I agree that the library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by the professor who supervised my dissertation work or, in his absence, by the chairperson of the department or the dean of the Graduate School. It is understood that any copying or publication or other use of this dissertation or part thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of North Dakota in any scholarly use which may be made of any material in my dissertation.

Signature Michelle D Dela

Date 11/29/2010

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

The author expresses thanks to Phil Padgett whose advice was sometimes accepted, as well as to the members of my committee, Dr. R. Landry, Dr. K. W. Gershman, Dr. D. Yearwood, and Dr. S. D. LeMire for encouraging me to develop this model.

To Dr. John L. V. Bobell

All Educators

ABSTRACT

A study was conducted to propose a research-based model for a longitudinal data research system that addressed recommendations from a synthesis of literature related to:

- needs reported by the U.S. Department of Education,
- the twelve mandatory elements that define federally approved state longitudinal data systems (SLDS),
- the constraints experienced by seven Midwestern states toward providing access to essential educational and employment data, and
- constraints reported by experts in data warehousing systems.

The review of literature investigated U.S. government legislation related to SLDS and protection of personally identifiable information, SLDS design and complexity, repurposing business data warehouse systems for educational outcomes research, and the use of longitudinal research systems for education and employment outcomes. The results were integrated with practitioner experience to derive design objectives and design elements for a model system optimized for longitudinal research. The resulting model incorporated a design-build engineering approach to achieve a cost effective, obsolescence-resistant, and scalable design. The software application has robust security features, is compatible with Macintosh and PC computers, and is capable of two-way live connections with industry standard database hardware and software. Design features included:

- An inverted formal planning process to connect decision makers and data users to the sources of data through development of local interactive research planning tools,
- a data processing module that replaced personally identifiable information with a system-generated code to support the use of de-identified disaggregate raw data across tables and agencies in all phases of data storage, retrieval, analysis, visualization, and reporting in compliance with restrictions on disclosure of personally identifiable information,
- functionality to support complex statistical analysis across data tables using knowledge discovery in databases and data mining techniques, and
- integrated training for users.

The longitudinal research database model demonstrates the result of a top down-bottom up design process which starts with defining strategic and operational planning goals and the data that must be collected and analyzed to support them. The process continues with analyzing and reporting data in a mathematically programmed, fully functional system operated by multiple level users that could be more effective and less costly than repurposed business data warehouse systems.

CHAPTER I

INTRODUCTION

Background

From January 2005 to January 2010, the U.S. government authorized approximately $4.8 billion in one-time grants to states for the development of state longitudinal data systems (SLDS). The purpose of these systems has been to support the analysis of education and employment outcomes to improve educational accountability and research processes from PK-12 through postsecondary, into the workforce, and across states. Despite generous federal grants, the success of a national system of interconnected state databases sharing educational and employment research data may be an unrealistic expectation. Typical SLDS include combinations of hardware and software re-purposed from complex business data warehouse applications that have a history of failures related to hardware, software, standards, organizational dynamics, and technical proficiency (Mullin & Lebesch, 2010).

A recent U.S. Department of Education (2010) longitudinal study of student data systems reported that 77.0% of the PK-12 districts surveyed were using data warehouses consisting of multiple software applications with over half of the districts reporting interoperability issues across disparate data systems. The same study indicated that over 80.0% of the districts reported difficulty using electronic systems to collect and analyze data that could support classroom teaching (U.S. Department of Education, 2010). In

other words, over 80.0% of the school districts reported barriers to the use of data for decision-making and school improvement. Leaders in the information technology (IT) infrastructure have reported success rates as low as 17.0% regarding typical business data warehouse systems repurposed for use in SLDS. Other leaders have reported that the major problems in IT projects result from data integration and information management issues, offering recommendations to separate data storage and retrieval functions managed by IT staff from data-driven research systems managed by teachers, administrators, and research practitioners. The U.S. government and the states that have received SLDS funding may be overlooking or bypassing provisions of existing laws related to the privacy of personal information (Bakst, 2009; Center on Law and Information Policy, 2009; Kline, 2010; Lederman, 2010, February). Finally, states who have elected to participate in the grant funded SLDS projects may have overlooked the sustainability issue related to a caution stated by the U.S. government that the billions of dollars expended for SLDS projects represent a one-time federally funded investment (White House, 2009).

Purpose

The situation described above constitutes a complex problem for all levels of education accountability and research across the United States. The purpose of this dissertation was to propose a model for a statewide longitudinal relational research data system that supports data-driven-decision making for the improvement of education and employment outcomes. The Simple Effective Actionable Knowledge Optimized for Research (SEAKOR) model fully complies with the twelve mandatory federal elements that define approved state longitudinal data systems. The SEAKOR model demonstrates a

cost effective obsolescence system that could be scaled at all levels of American education systems in support of school improvement, accountability, and reform. The SEAKOR model supports data-driven decisions related to formative and summative assessment of teaching and learning, as well as the evaluation of curriculum, programs, administration, staffing, and teacher education. The SEAKOR model has related application to institutional accreditation, strategic planning, and policy-making.

Rationale for the SEAKOR Model

The SEAKOR model responds to the need for a model education data system reported by the U.S. Department of Education (2010), as well as needs and issues reported by experts in data warehousing systems and experts in education regarding the responsible use of data for education and employment outcomes. The SEAKOR model addresses the U.S. Department of Education finding that most education data warehousing systems are "...so complex and poorly aligned that their use by school staff was not feasible" (p. 2).

The SEAKOR model also responds to the U.S. Department of Education finding that the complexity of existing multiple data storage and retrieval products that make up educational data systems severely constrain research, as well as the finding that system interoperability success is less than 40.0%. Given credible reports that data warehousing system complexity is related to statistical failure rates as high as 83.3% (Charette, 2006), the SEAKOR best practices model also addresses Laird's (2008) view that a complex statistically vulnerable "...data warehouse system is not needed to share records... states can share records now by making small adjustments..." (p.5).

The proposed SEAKOR model demonstrates a simplified, cost effective, obsolescent resistant system that could be scaled up or down for implementation at all levels of American education systems to support strategic planning, accountability, school improvement, classroom learning, and reform. The SEAKOR model is designed to fully comply with the twelve mandatory federal elements that define approved SLDS (America COMPETES Act, 2007). I also attempted to respond to the specific concerns of Illinois, Indiana, Iowa, Michigan, Minnesota, Ohio, and Wisconsin regarding the use of data systems to improve education processes and outcomes (McDonald, Andal, Brown, & Schneider, 2007). I attempted to respond to unique local or regional needs by designing into the model system the ability to use data translation and interpretation methods that would accommodate differences in data systems across states without the need to resort to expensive changes to existing systems to support interconnectivity.

The SEAKOR model described herein is designed to avoid or minimize the effects of barriers to success of business data warehouse systems being repurposed for use in SLDS. Arguably, because of the realities of economic accountability, American business demands higher performance in information technology systems than does American education. On the other hand, data collection, storage, retrieval, analysis, and reporting are, in aggregate, significantly more complex processes (Smith, 2004) in education and employment research than in business where systems design is more focused and linear (U.S. Department of Education, Office of Planning, Evaluation, & Policy Development, 2010).

Rationale for the Study

Data management used in business and education today generally consists of multiple layers of commercial off-the-shelf hardware and software that support discrete data management processes. Such data systems generally require the need for additional layers of software to compensate for performance constraints imposed by the initial design layers, a process that introduces issues of complexity, sustainability, obsolescence, and performance (Alves & Finkelstein, 2002; Basken, 2010; Khabaza, 2009; Tiwana, 2002). An example of complexity comes from the state of California that has in excess of 125 separate educational data collection efforts ongoing on a concurrent basis (Children Now, 2009). These incompatibility issues could limit the effectiveness of a national system that depends on complete, valid, and reliable data in local and state systems (McDonald, Andal, Brown, & Schneider, 2007; Prescott & Ewell, 2009).

Method

The method consisted of six general steps:

1. Reviews of U.S. government legislation related to SLDS grants were conducted to determine SLDS grant eligibility and performance requirements.
2. Reviews of the experiences of seven Midwestern states that received SLDS grants were reviewed to understand progress and impediments to progress toward achieving SLDS grant requirements.
3. A review of the U.S. Department of Education report on the use of education data at the local level was conducted to understand progress and impediments toward reaching U.S. Department of Education goals for data-driven decision-making practices.

4. Information gathered in steps one, two, and three were studied to determine areas and topics for additional study in relation to barriers to accomplishing U.S. government goals for SLDS. These areas included the complexity of SLDS, barriers to success of SLDS, protection of personally identifiable information, U.S. government privacy legislation, reports of failures to protect personally identifiable information, unauthorized disclosure of personal information, risks associated with disclosing personally identifiable information, business IT project performance, business IT obsolescence, evolution of data processing systems, limitations of business data warehouse architecture and designs, SLDS design requirements, SLDS project funding, a need for a model longitudinal research system for educational and employment outcomes, best practices for longitudinal research data system engineering and design, and relating the need for a model research data system to failures in business data warehouse systems.
5. Determining the capabilities and features necessary in a model longitudinal research data system that would address the barriers to SLDS success identified in previous steps.
6. Designing the SEAKOR model longitudinal research data system.

CHAPTER II

LEGISLATION

Arguably, little has changed since the Director of the Eight-year Study wrote that he was amazed that schools and colleges knew little of their work and that they seldom attempted to discover what changes occur in students as a result of education (Aikin, 1942). It would seem that Aikin was describing the need for an effective research data system that could evaluate education and employment outcomes, a need that has yet to be fulfilled almost 70 years later, in spite of massive amounts of federal grant funding. For example, from January 2005 to December 2009, the U.S. government authorized approximately $4.8 billion in grant funding to support several related programs that promote P-12 state longitudinal data systems (SLDS) in support of education accountability and research (Mullin & Lebesch, 2010; U.S Department of Education, Office of Planning, Evaluation, & Policy Development, 2010). According to Mullin and Lebesch (2010), four major legislative acts have shaped the requirements for accepting federal SLDS funding.

Education Technical Assistance Act of 2002

The Education Technical Assistance Act (2002) provided $265 million to 41 states from January 2006 to January 2009. The purpose of this act was to encourage the development of longitudinal data systems to study education outcomes of students from Pre-kindergarten through secondary.

America COMPETES Act of 2007

The America Creating Opportunities to Meaningfully Promote Excellence in Technology, Education, and Science Act (America COMPETES Act, 2007) built upon the state longitudinal database system initiative provided in the Education Technical Assistance Act (2002) by promoting increased accountability regarding the preparation of K-12 students for higher education, the workforce, and membership in the U.S. military. The America COMPETES Act legislation specified disaggregate data elements to be included in U.S. government funded state longitudinal data systems.

1. Unique state specific student identification code,
2. Disaggregate student data related to demographics, programs, and enrollments, including information regarding entry, exit, transfer, dropout, and completions;
3. The ability to link PK-12 data to higher education system data,
4. Information regarding longitudinal data system data quality, validity, and reliability;
5. For PK-12, disaggregate annual test results, data on students not tested, the ability to match teachers with students, transcripts, and college readiness test results;
6. For postsecondary education, disaggregate data related to PK-12 transition to postsecondary education, including remedial course data.

While appropriations of funds to support the America COMPETES Act were not authorized, the data conditions for states have not been rescinded and continue as requirements in subsequent legislation related to SLDS. The stated intent of the U.S. Government investment in these related programs has been to encourage innovation and

education reform through the alignment of data from multiple state agencies. However, the U.S. government appears to be using the SLDS grant initiative to fund an extensive interconnected national system of K-12 state longitudinal data systems (Kline, 2010; Lederman, 2010, May).

Higher Education Opportunity Act of 2008

The Higher Education Opportunity Act (HEOA, 2008) prohibited the development of a U.S. government federal level data system to track disaggregated student data. However, the language of the act did not include states in the data system prohibition. The HEOA authorized a pilot program for multiple states to test the development of state-level disaggregated postsecondary student data systems. Funds for the pilot program were not authorized. Though unstated, the U.S. government's purpose for developing state data systems into an interconnected and decentralized national system would appear to be that of accomplishing the goals of a centralized federal education and workforce system without violating the carefully worded prohibitive language in the HEOA (Bakst, 2009; Basken, 2010; Mullin & Lebesch, 2010).

American Recovery and Reinvestment Act of 2009

The American Recovery and Reinvestment Act (ARRA, 2009) included multiple initiatives to support economic stimulus and creation of jobs. Regarding education, the ARRA provided funding to support at-risk education jobs, as well as funding for school modernization projects and tuition tax credits. Of the $141.4 billion allocated for education by the ARRA, $5 billion was assigned to promote robust data systems (Department of Education, 2009a). According to the Data Quality Campaign (2009) the ARRA expanded support for SLDS with language that explains the intent of congress to

build systems that establish Pre-K to college and career data systems that track student progress, including transition from secondary to postsecondary and enrollment in remedial coursework. Imbedded in the ARRA legislation was a controversial competitive grant fund of $4.35 billion known as the Race to the Top. Race to the Top competitive U.S. government grant criteria were weighted to favor states based on the extent of P-12 SLDS capabilities to access and use state education data to improve instruction (U.S. Department of Education, 2009b). A condition of acceptance by recipient states was the burden of compliance with the specific provisions regarding specific disaggregate data elements included in the America COMPETES Act.

Student Aid and Fiscal Responsibility Act of 2009

The Student Aid and Fiscal Responsibility Act (2009) was incorporated as a rider on the Health Care and Education Reconciliation Act (2010) signed into law on March 30, 2010. The act replaced the student loan system with a U.S. government direct loan program and revised legislation regarding financial aid to students. An earlier version of the Student Aid and Fiscal Responsibility Act written in the House of Representatives included Section 505, National Activities, intended to establish a Learning and Earning Research Center under the Director of the Institute of Education Sciences (Student Aid and Fiscal Responsibility Act, 2010). The Learning and Earning Research Center would have been authorized to develop data elements, definitions, and data sharing protocols to link postsecondary data systems across states participating in the SLDS grant funded activities. The Secretary of Education would have been authorized to award grants to states to develop and implement SLDS that would share disaggregate student data from community colleges linked to elementary and secondary education and workforce data

systems using funding appropriated in the ARRA (2009). While the language of the proposed amendment prohibited the disclosure of personally identifiable information, complying with the data sharing requirements would appear to require states to ignore details in existing federal laws (Krigman, 2009; Lederman, 2009). The proposed House amendment failed to receive joint House-Senate approval and did not become part of public law.

Summary

This section has summarized the four U.S. government legislative acts that fund and implement a national system of SLDS that could be the basis for fulfilling a perceived, but unapproved, U.S. government need for a centralized federal-level student information system (Bakst, 2009; Higher Education Opportunity Act, 2008). Chapter III will discuss some of the issues and constraints associated with typical SLDS designed funded by state and U.S. government funds.

CHAPTER III

STATE LONGITUDINAL DATA SYSTEMS

Despite generous U.S. grant funding, creating a national system of interconnected state longitudinal data systems (SLDS) for education and employment outcomes research may be an unrealistic expectation (Basken, 2010). Hurdles to linking and sharing education and employment records within and across states would seem to exist in several areas. First, while complying with legislation related to the protection of personally identifiable information, each state data system must be designed and built so that a variety of data can be collected and stored across education and other state agencies, including employment. Second, each state data system must be designed to relate data on individuals from a variety of sources into records that capture and display information for data analysis.

Complexity of State Longitudinal Data Systems

One purpose of an education longitudinal state data system is to consider all education and employment records related to an individual within a single state system for statistical analysis. Individual education and employment databases managed by separate state agencies could include millions of records that must be electronically linked to support complex analysis. From a research perspective, it may be difficult for an individual not routinely involved in research regarding the outcomes of education and employment to comprehend the complexity of multiple database systems that are

electronically connected to evaluate data across connected database systems as if the data were contained within one data system accessed by one computer. For example, one of the typical purposes of a state longitudinal database system is to evaluate the outcomes of higher education that could include analyzing graduate records of programs of study by classification of instructional program, employment following graduation, re-enrollment in subsequent education programs, or a combination of these three major sources of data contained in multiple, but separate, databases. The complexity is compounded when individual graduates may have completed multiple degree programs involving multiple institutions within the same academic year, as well as situations in which individuals may have multiple careers with multiple strings of income, and perhaps, participation in multiple social programs. An example of the complexity is illustrated in Figure 3.

Individual school districts or state agencies may connect their computer workstations to one or more servers dedicated to a single purpose or process. These servers may be integrated into one server system known as a data mart. A data mart is a database or data table containing data limited to a specific purpose or subject. An example could be enrollment records. A state that has 92 school districts and eight state agencies involved in collecting, storing, and sharing data may have 100 data marts, at least one for each school district and state agency. Multiple data systems introduce the possibility of compatibility barriers that could affect the success of an interconnected system. Ideally, the state data marts are successfully connected to a larger server system that is specially configured to process and report aggregate data. This larger server system is typically known as a data warehouse. A data warehouse contains data from various sources that is pre-processed and validated before being used for analytical

procedure. While the primary purpose of a data mart is data storage and summarization, the purpose of a data warehouse is data analysis and aggregate reporting (Cios, Pedrycz, Swiniarski, & Kurgan, 2007). For example, a data mart may capture data from a school information system that provides real-time accounting of daily school functions. In contrast, a data warehouse provides access to historic data but may not be designed for immediate access to new data pending processing (Wayman, 2005).

The projected decentralized system of 50 state longitudinal data systems connected into a national network must be designed and built in such a way that these data systems can communicate information from one state to another state as individual school districts and state agency data marts are intended to communicate with state data warehouses. These systems may be visualized as linear networks of sequential components connected into layers with each layer adding a new dimension to the existing complexity as shown in Figure 1. A malfunction in a single sequentially linked component, such as an identifier shared among matched records, could significantly affect the reliability of the entire sequentially connected system.

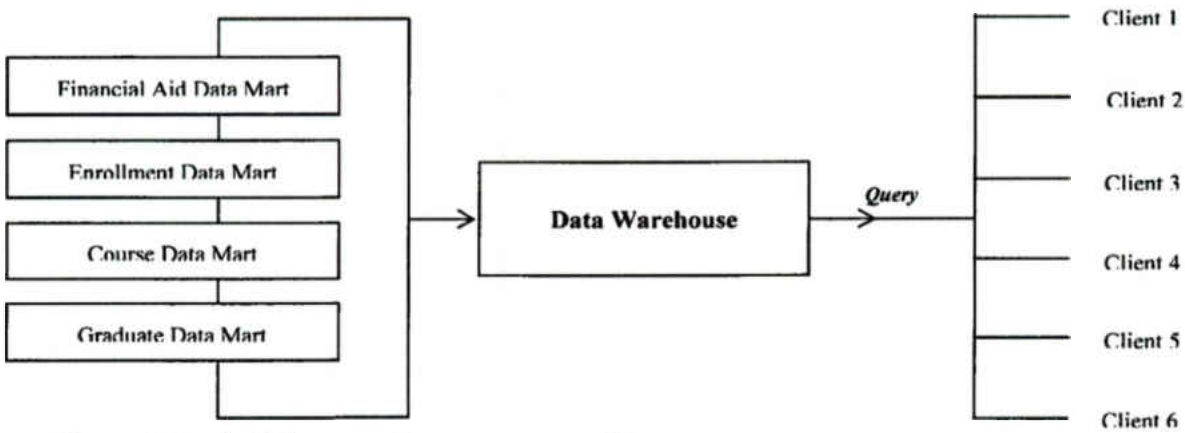

Figure 1. Typical data warehouse system architecture.

New data are generated as event-related longitudinal transactions that may include records of different homes, schools, teachers, courses, tests, programs, employers, and sources of income across multiple agencies and perhaps in multiple communities within a state. The complexity of relating records for research purposes are compounded when multiple states become involved, such as following the outcomes of individuals who could be schooled in one or more states and relocate to different states for employment. Typical procedures, based on the limitations of technology, attempt to aggregate all these data related to one individual into a single record for statistical analysis. This process of compressing data into aggregate records destroys the ability to conduct the level of research that state and federally funded systems are intended to support. Therefore, to achieve research goals, data systems must be designed to systematically process and analyze disaggregate data from multiple sources within the restrictions that apply to the protection of personally identifiable information. A major source of stress to states is balancing compliance with one set of rules without violating seemingly conflicting rules, a situation that requires reconciling a dichotomy with regard to the need for access to disaggregate individual data which must be retained to support research that bridges state agencies and multiple years.

The research issues related to multiple records of individuals contained in multiple databases across multiple systems are, in some cases, negatively influenced by the need for a patchwork of multiple hardware and software combinations at each level that serve the data processes of collection, retrieval, pre-processing, analysis, visualization, and reporting. In other words, data from individual state data systems may

not be compatible with data developed by other states' systems due to differences in the way different organizations define the data contained in database fields and other information technology design factors. Lack of coordination and compatibility compounds the vulnerability while adding to the complexity of managing such systems. For instance, a recent U.S. Department of Education (2010) longitudinal study of student data systems reported that 77% of the P-12 districts surveyed had data warehouses consisting of multiple software applications. Over 60% of the districts surveyed reported a lack of interoperability across multiple data systems that constituted a barrier to the use of data for decision-making and school improvement. School districts reported frustration at the inability to link their multiple district data systems. The problem is compounded when district data systems are interconnected into a state education agency level system and other state research systems. For instance, the state of California may have more than 125 concurrent data collection efforts (Children Now, 2009). These incompatibility issues could limit the effectiveness of a national system that depends on complete, valid, and reliable data in local and state systems (McDonald, Andal, Brown, & Schneider, 2007; Prescott & Ewell, 2009). According to Kowalski, Lasley, and Mohaney (2008), systems must be designed to use exactly the same metrics, measures, and procedures. Alternatively, systems must be designed with an intermediate translation capability that allows different systems to communicate through interpretation. Translation and interpretation features could also mitigate issues involved with comparing standardized test results across systems and states, a benefit that could avoid the "unforeseen negative consequences" (Prescott & Ewell, 2009, p 1) of compatibility.

Barriers to Success of State Longitudinal Data Systems

The National Center for Education Achievement, reported additional barriers that include the lack of resources, lack of common student identifiers, lack of coordination, and threats to student privacy (Laird, 2008). According to the Washington State Attorney General:

> The pervasive use of technologies in our everyday lives and in our work gives rise to the potential compromise of personal data privacy if appropriate care is not taken to protect personal information. Unfortunately, many individuals are unaware of privacy laws and data protection and how they can help ensure the privacy of their personal information online (McKenna, 2009; ¶ 2).

Privacy Legislation

The most significant student privacy issues are restrictions on the disclosure and use of personally identifiable information protected under Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2002, Family Educational Rights and Privacy Act (FERPA) of 1974, the Health Insurance Portability and Accountability Act (HIPAA) of 1996, and the Protection of Pupil Rights Amendment (PPRA) of 1978. Personally identifiable employment confidentiality is governed under CIPSEA. Disclosure of protected employment data is a felony under 31 U.S.C. 1104(d); 44 U.S.C. 3504 with penalties that include five years of imprisonment and up to a $250,000 fine. Federal statute 20 U.S.C. § 1232g; 34 CFR Part 99 or FERPA protects an individual's right to privacy regarding educational records. Violations of FERPA may result in the loss of federal funds to the offending institution; however, the penalty has not yet been imposed (Mills, 2005). Personally identifiable health information is governed under PL

104-191 or HIPAA. Violations of HIPAA may result in fines per violation and could send a violator to criminal court where penalties range up to $250,000 and ten years imprisonment (Shiley, 2003). Violations of PPRA or 20 U.S.C. § 1232h; 34 CFR Part 98 include warnings, withholding federal funding, and/or termination of federal funding.

Failures to Protect Personally Identifiable Information

This legislation was established to ensure the protection of individual privacy data. Arguably, the U.S. government would be expected to enforce stricter safeguards for protection of privacy information than an interconnected system of state databases. However, the federal government has experienced significant breaches in security. According to the U.S. Government Accountability Office (2008), from 2003 to 2006, 19 U.S. government departments and agencies reported at least one breach of personally identifiable information that could expose multiple individuals to identity theft. The U.S. Government Accountability Office reported a significant increase in incidents from 2005 with 3,469 incident reports to 2006 with 5,146 incident reports. The estimated cost of identity theft in 2006 within U.S. organizations was estimated to be $49.3 billion.

Pressures to Disregard Personal Privacy

Since 2005, the federal government and states that are receiving federal funds to implement a national system of interconnected state databases may be overlooking or bypassing existing laws related to personal privacy protection (Lederman, 2010, February). In February 2010, Representative Kline (2010), Senior Republican Member of the Committee on Education and Labor wrote a letter to the Secretary, U.S. Department of Education, expressing significant concern that the U.S. Department of Education was placing personally identifiable student information at-risk with requirements levied as

conditions for accepting U.S. government funding for SLDS. Representative Kline indicated concern that the administration's policies did not appear to comply with the privacy requirements of FERPA. Representative Kline noted that congress had not authorized the Department of Education to facilitate the creation of a national student database and reminded the Secretary Duncan of prohibitive language of legislation, including the Elementary and Secondary Education Act and the Higher Education Opportunity Act. "As important as effective research is to successful education reform, student privacy protections must not be forced to take a backseat" (¶ 7).

Bakst (2009), an attorney and President of the Council on Law in Higher Education noted that the U.S. Federal Trade Commission, an entity of the federal government, has established five principles for the use of privacy information that are, so far, absent in discussions of the movement to undermine the protections currently afforded students under FERPA. The five principles established by the U.S. Federal Trade Commission (2007) include:

1. Notice/awareness,
2. Choice/consent,
3. Access/participation,
4. Integrity/security, and
5. Enforcement/redress.

These principles specifically relate to adults and parents regarding the status of children as a special vulnerable group. In other words, under the U.S. Federal Trade Commission's Principles regarding privacy, parents have the means to control collection and use of their children's personal information.

The literature researched for this dissertation includes no evidence of principles or guarantees regarding privacy for students beyond those currently established in FERPA, HIPAA, CIPSEA, and PPRA regarding development and implementation of SLDS that use personally identifiable data. For example, it would not appear that parents, under the envisioned policies for SLDS, would have the same rights to protect their children within the SLDS as they would have to protect their children against telemarketers under the five principles established by the U.S. federal Trade Commission.

The Fordham Center on Law and Information policy (Center on Law and Information Policy, 2009) reviewed student privacy protection status and issues across the 50 states. The purpose of the Fordham study was to investigate what type of data was being collected on children and to determine if the processes in use protected children legally and technically from data misuse and improper data release including data breaches. The Fordham study found that privacy protection in longitudinal databases to be lacking in the majority of states. For example, 32.0% of the states' data warehouses contained student social security numbers. Additionally, 22.0% of the states' data systems reported children's pregnancies and 46.0% of the states tracked mental health and jail sentences as part of children's educational records. The Fordham study also found that several states use out-of-state contractors to warehouse data without any protections for privacy established in contracts with vendors. Fordham's findings reflect confusing signals with regard to the need for protecting privacy and, perhaps, lack of leadership at levels of the federal and state governments responsible for implementing SLDS. Generally, the Fordham study found that the flow of information from educational

agencies to state departments of education was not in compliance with the provisions of FERPA.

Unauthorized Disclosures of Personal Information

In April 2007, the Washington Post reported unauthorized intrusions of federal databases operated by the U.S. Department of Education that contained confidential information on approximately 60 million recipients of student loans. The database was "repeatedly searched... in ways that violate federal rules, raising alarms about abuse of privacy..." (Paley, 2007; ¶ 1). The abuses were disclosed as the $85 billion per year student loan industry was under investigation for lack of oversight. Approximately 29,000 authorized financial aid administrators have access to the student loan database along with approximately 7,500 loan company employees who were found to be using student information in marketing campaigns. Financial aid directors reported the abuse to the federal government with stories of students who had previously qualified for direct loans with the federal government receiving up to six solicitations per day from private loan companies who had access to the database.

In an article critical of the U.S. government's lack of student privacy enforcement, the Council on Law in Higher Education, Bakst (2009) contrasted the prohibitive language of the Higher Education Opportunity Act of 2008 regarding a centralized, federal student data tracking system with the U.S. government's continued funding support to assist states in creating a more vulnerable national system of state databases to function as a surrogate federal system. Bakst argued that large state data systems place privacy at-risk without evidence that large data systems could be linked to higher performing education systems. Bakst challenged the U.S. government to explain

why such large interconnected systems were so important as to justify the risk of breaching the privacy of students.

It is interesting to note that the most recent Race to the Top legislation fails to address individual privacy protection and also fails to predict the educational benefits to be derived from implementing SLDS (U.S. Department of Education, 2009b). "We are rushing, again, into school reform initiatives with billions of dollars without much evidence that where we are headed is the right direction, and, in some cases, with evidence that it is clearly the wrong one [sic]" (Strauss, 2010; ¶ 8).

The Privacy Rights Clearinghouse (PRC, 2010) maintains a searchable database of reports of breached personal privacy information. The PRC database can be filtered by unintended disclosure; hacking or malware; unintentional breach by employee or contractor; lost, discarded, or stolen records; lost, discarded, or stolen electronic devices that store privacy data; lost, discarded, or stolen stationary electronic devices, such as computers and servers; and unknown sources. For the years 2006 to 2010, the PRC reported 372 breaches in educational systems representing 5,624,105 privacy records (PRC, 2010). Increasing the number of employees, researchers, and policymakers who have access to privacy data and increasing the number of systems that store or display privacy data, or increasing the number of ways that privacy data can be accessed would seem to increase the risks of intentional or accidental breaches of security. Basken (2010), cautions that those in favor of student record databases concede that the proposed national data system with interconnected state systems has unaligned standards for data collection, analysis, and data protection. Basken argues that the proposed decentralized system poses a greater risk to overall privacy than a centralized national data system.

Risks Associated with Disclosing Personal Information

In addition to security breaches and inappropriate use of privacy information, the proliferation of properly-stored and appropriately-used information may put individuals at- risk if the information used and shared across systems contains incorrect information. Properly-stored and used incorrect information could endanger an individual's education, employment, social security, social benefits, legal status, and/or financial status. Credit agencies are required to disclose records used to reach an unfavorable decision for a loan request. A search of literature does not indicate similar protection for inaccurate information stored on individuals whose privacy records are being shared across multiple systems and states.

The National Student Clearinghouse (NSC) stores student records submitted by 3,300 institutions of higher education for data sharing and research. The NSC website (2010) claims credit for storing the records of 92.0% of U.S. college students. These data contain social security numbers, birth dates, full name, enrollments, completions and other such information. Some states consider the NSC a major source of research data, as well as a model for a retrievable student data sharing system. In 2009, the NSC initiated a service titled Secondary Education Research Initiative that was partially funded by the Gates Foundation to expand an existing service known as StudentTracker for high schools. The StudentTracker service allows administrators of high school districts to access the clearinghouse records of approximately 100 million students. The NSC website (2010) states that the participating colleges and universities have authorized the clearinghouse to provide postsecondary education data to high schools and high school districts to support educational program improvement.

The NSC was established with the support of student loan grantors to assist with verification of student loan applicants, and its role has expanded to serve businesses with verifying the higher education credentials of job applicants. Higher education institutions cooperate with the NSC because clearinghouse data can be used to increase their published graduation rates as a result of tracking down former students who completed their education at other institutions and perhaps in other states (Basken, 2010).

However, while the NSC may be considered by some to be a viable model for sharing student data (Basken, 2010), the clearinghouse is noted for inaccuracies in the data that it stores and shares. The NSC does not provide a remedy for individuals to correct their personal records even though students whose records are being stored may not have authorized the NSC to store or share their personal information. Students may be unaware that their records have been disclosed. If inaccurate student data stored in the NSC is shared with potential employers about education credentials, the incorrect data could prevent a qualified individual from securing a professional position.

If inaccurate student data are shared with financial institutions or educational institutions regarding educational qualification for student loans, the incorrect data could prevent a favorable decision on a loan request. Arguably, the NSC data sharing model could pose greater risks to student privacy than SLDS systems with procedures and constraints that could harm students through properly-stored but inaccurate data that is difficult for students to correct.

Without strict standards and consensus for data accuracy and interconnectivity of data systems across states, and without standards or consistency with regard to protection of privacy information, it is reasonable to assume that a significant percentage of the

millions of student records and a significant amount of the billions of dollars allocated for SLDS are at-risk.

Summary

This chapter discussed SLDS issues related to system complexity and barriers to success, such as resource limitations and the lack of standardized student identifiers necessary to match student records across linked databases. Also discussed were issues related to individual rights to privacy, including legislation, pressures to disregard privacy laws, unauthorized disclosure of personally identifiably information, and other risks and consequences. Chapter IV discusses the use of business database system hardware and software for statewide longitudinal database systems, a practice that could increase the overall complexity and cost of educational data systems while adding additional barriers to SLDS success related to the statistical vulnerabilities of using business data systems for educational and employment research.

CHAPTER IV

REPURPOSING BUSINESS DATA WAREHOUSE SYSTEMS IN EDUCATIONAL OUTCOMES RESEARCH

Education database systems and business database systems frequently use the same hardware and software from the same manufacturers for similar purposes. Therefore, education and employment database systems may inherit design and performance characteristics from business systems that share the same hardware and software. Arguably, economic accountability considerations in business settings may result in the demand for higher database performance and reliability than exists in American education. On the other hand, education and employment data collection, storage, retrieval, analysis, and reporting, in aggregate, are significantly more complex processes than are found in business where system design may be more repetitive, focused, and linear (Cios, Pedrycz, Swiniarski, & Kurgan, 2007; Palaich, Good, & van der Ploeg, 2004), such as in recording and aggregating financial transactions and changes in inventory.

The spending for all U.S. information technology (IT) goods and services for 2010 is estimated to increase 9.9% to $564 billion compared to a 7.8% increase worldwide for the same period (Kanaracus, 2010). Business related IT represents approximately 50% of all business equipment spending in the U.S. However, the literature suggests an annual failure rate of 83.3% of business IT related projects in 2002 (Tiwana, 2002). According to Charette (2006), the president of an IT risk

management consulting firm, the cost associated with IT project failures in the U.S. was estimated to be $60 to $75 billion per year.

Bodamer (2010) indicated that the situation involving government projects is particularly acute. In April 2005, the U.S. Federal Bureau of Investigation (FBI) abandoned a $170 million longitudinal database including 700,000 lines of unusable code (Goldstein, 2005). The U.S. Department of Justice Inspector General (FBI, 2005) released an 81-page audit of the FBI project in 2005 that described factors that contributed to the FBI project failure. These factors included 800 pages of poorly defined system requirements, multiple changes in design, unrealistic project schedules, and the lack of an effective procurement plan. Contributing to the cost of failure was the cost-plus-award procurement contract that essentially allowed for uncontrollable growth without accountability. In March 2004 a litigation arbitrator found that of 59 specific problems, 19 were related to FBI initiated changes to project requirements, and 40 were related to contractor errors (Goldstein, 2005). Charette (2010) reported that the redesigned replacement for the FBI database project was currently behind schedule with an estimated cost of $425 million based on a 2009 completion date. However, the cost has grown to almost $557 million with a revised earliest completion date of 2011. The purpose of the FBI virtual case file project was to consolidate all of the FBI historical investigative case research into an easily accessible data system for the use of 12,400 individual agents assigned to 56 field offices, 400 satellite offices, and 51 legal attaché offices in U.S. embassies (Goldstein, 2005), a situation similar to creating a national system of interconnected educational data systems. The scope and complexity of the FBI longitudinal data warehouse project with 700,000 lines of code with multiple changes in

design at the time of abandonment would seem to be a candidate for the non-linear interaction database performance issues described by Agrawal (2005).

Agrawal (2005), a doctor of engineering manufacturing management, conducted research in simulating performance issues in data warehouse design. Agrawal reported that the complexity of data warehouses is related to interactions among non-linear components, in that a small change in one component could introduce dramatic changes elsewhere in the system with unpredictable results. Additionally, Agrawal noted that data warehouses are dependent upon other systems for data and that response of the entire system is often difficult to predict. The sometimes extreme sensitivity to small changes in a data warehouse system indicated to Agrawal that the application of Chaos Theory could be used to study the erratic performance that he discovered during his data warehouse simulation research. A system is considered chaotic when its sensitivity depends upon initial conditions. In order for a deterministic system to be chaotic, the system must be non-linear. Controlling chaos is a process in which a very small disturbance is applied to realize a desirable behavior (Boccaletti, Grebogi, Lai, Mancini, & Maza, 2000).

Business Information Technology Project Performance

IT project management consists of five phases: Initiation, planning, execution, control, and closure. According to Evans (2005), most problems with IT development result from the separation of project initiation from project execution. During the selection of bidders, vendors, or advisors, projections of costs may be minimized while projections of outcomes may be overstated to engage support for a project. This occurrence could be particularly true in public sector IT projects where, arguably, accountability to taxpayers is less rigorous than accountability to business owners in an

environment where litigation could be a consequence of poor performance by project managers and vendors. Understating or underestimating costs may lead to inadequate start-up budgets that, in turn, may lead to various maneuvers to increase productivity, adjust outcome expectations, or take risky shortcuts in project validation and testing. These activities directly increase the statistical probability of IT project failure (Charette, 2006).

In terms of performance, a study of 8,000 information technology projects started by 400 U.S. businesses determined a success rate of 16.3% while 20.0% were never completed and 41.3% were completed over-budget and/or were completed after significant delays and/or failed to accomplish design goals (Tiwana, 2002). In a study conducted from 1994 to 2000 involving 30,000 information technology projects, project success ranged from 17.0% to 28.0%. Abandoned projects ranged from 20.0% to 51.0%, depending on the year of study. Of the projects studied, 32.0% to 52.0% were completed over-budget and/or were completed after significant delays and/or failed to accomplish design goals (Nemati, Steiger, Iyer, & Herschel, 2002). While a typical data warehousing project may cost in excess of $1 million each, the failure rates exceed 50% (Charette, 2006). The annual cost of failure in information technology projects to American business exceeds $78 billion in development costs with an additional $22 billion in cost overruns. An internationally recognized authority in risk management IT systems engineering, and large-scale, software-intensive data systems indicates the $78 billion in annual losses due to IT project failure does not include the costs of projects that simply exceed their budgets, the costs associated with projects that are completed late, the additional costs associated with re-designing a project once abandoned, or the costs of

marginally performing projects that require unplanned resources for continuous attention (Charett, 2006).

Most IT project failures are the result of multiple factors that can be summarized as a combination of flawed project management decisions, flawed business decisions, and flawed technical decisions. Charette's (2006) research reported the 12 most common factors in the failure of IT projects to be:

1. Unrealistic or unarticulated project goals,
2. Inaccurate estimates of needed resources,
3. Badly defined system requirements,
4. Poor reporting of the project's status,
5. Unmanaged risks,
6. Poor communication among customers, developers, and users,
7. Use of immature technology,
8. Inability to handle the project's complexity,
9. Sloppy development practices,
10. Poor project management,
11. Stakeholder politics and,
12. Commercial pressures.

Since 2004, the Standish Group has noted an improvement in statistical IT project success from approximately 16.3% to approximately 34.0%. One reason for improvement in IT project success may be related to Bodamer's (2010) observation of a possible cultural shift in which project outcomes are administratively reduced in order to adhere to original costs and deadlines. This explanation would seem to be a possible method to

claim partial success for an otherwise failed IT project if it was found that the original project needs could not be accommodated within budget.

Another reason for the improvement in statistical IT project success may be a change in project design that focuses on developing and implementing smaller scale projects using an iterative planning process in contrast to methods that required complete project definition in the early phases of planning. A more successful approach to IT project planning could be investing more resources in the beginning to design IT projects that fulfill needs rather than depending on technology alone to solve operational and compatibility problems after project completion (Weinberger, 2004). According to Smith (2004), business leaders may view IT project failure as an unfortunate but necessary part of achieving competitive goals and are willing to risk a series of failures in order to reap the benefits of an innovative business or service that provides a company a competitive advantage. In contrast, "In education, unfortunately, failure is failure both inside and outside the classroom" (Smith, 2004, p. 96).

Business Information Technology Obsolescence

A factor in the long-term cost of IT projects is the obsolescence-prone nature of commercial off-the-shelf software. The literature describes three categories of obsolescence associated with IT systems that include data warehousing systems, hardware, software, and personnel knowledge and skill (Schneider, 2005). Obsolescence of hardware would appear to have the least impact on operations and sustainability, in that obsolete hardware can probably be replaced by more technologically advanced hardware. However, software obsolescence can obsolete hardware if software upgrades cannot interface correctly with the original hardware. If software obsolescence drives

hardware obsolescence, the consequences may include the need for a new system (Sandborn, 2007). A continuing concern with commercial off-the-shelf software is the cascading effect of dealing with software that is no longer supported due to planned obsolescence or obsolescence forced on an owner when a hardware upgrade is discovered to be incompatible with the corresponding software. A related situation occurs when a business takes over or consolidation results in a product's termination. According to Merola (2006), successful software vendors may render their own products obsolete by adding features in other products that create the need to upgrade the artificially obsolete products in order to support increased profits while taking credit for upgraded functionality. Schneider (2005) reports that a side effect of obsolescence in hardware or software could be creating an IT staff and user base that may have obsolete skill sets in coping with new technology. These factors may drive a need for training and/or reorganization to integrate replacement hardware and/or software into the organization's overall work flow. The risk of obsolescence may be directly related to the complexity of IT systems. In other words, a system with multiple servers and multiple software applications that support several different processes would pose a greater risk of obsolescence than a more austere system. Focused data management products generally outperform one-size-fits-all general-purpose products in cost, speed, and effectiveness. While less complex, such systems are typically more comprehensive (Monash, 2006).

Related to the effects of obsolescence in software is the obsolescence in data system architecture, as well as the factors that contribute to the limitations and failures in data warehouse systems is the effect of using aging technology. Stonebraker is currently the cofounder and Chief Technology Officer of Vertica Systems, a developer of an

emerging class of database systems. Stonebraker was formerly a professor of computer science at the University of California at Berkeley and the Massachusetts Institute of Technology. In IT history, Stonebraker was the cofounder of the architecture built into most of the relational databases used in data warehousing systems since 1970, including Oracle, SQL Server, Informix, and Sybase. Stonebraker is a recipient of international awards in computer science, including an international award for his work in designing the architecture of the listed database systems (Lai, 2007). Stonebraker (2007b) considers most of the systems used currently for data warehouse applications to be obsolete "legacy systems" (Stonebraker, 2007b, ¶ 3) in relation to systems designed since 1997.

Evolution of Data Processing Systems

Computing has steadily evolved since the Electronic Discrete Variable Automatic Computer (EDVAC) in 1949. EDVAC was the first computer with a program stored in the computer's memory. In other words, EDVAC could execute different routines by changing the content of the computer's memory, a process that previously required rewiring a computer. In EDVAC, the component that executed instructions sent to it from the computer's internal memory was called a processor or central processing unit (CPU). EDVAC was not practical. It averaged approximately eight hours between failures because early computers were constructed with electrical relay and vacuum tube technology that was not durable in computer applications.

During the 1950s and 1960s, transistor technology replaced earlier CPU construction methods. Transistors are tiny functional and more durable versions of vacuum tubes. A modern microprocessor is a memory chip that may include millions of miniature transistors and capacitors that are paired to create millions of memory cells,

each of which represents one bit of data. A bit of data is information that is stored in the form of a zero or one, essentially on or off. Sets of memory cells are used to create instructions in binary logic that computers understand. Machine language is a code that consists of a collection of binary bits that a computer reads and interprets to execute commands and is the only language a computer is capable of understanding. A variety of computer programming languages used more recently essentially send groups of commands to a computer's CPU using forms of notation that can be understood by human programmers. The programming language translates commands into machine language so that a computer can understand the instructions in machine language. Examples of common programming languages include Java, C++, Basic, Cobol, and Fortran.

Modern CPUs store programs, of which four program steps are common. The four program steps are fetch, decode, execute, and writeback. Fetch involves retrieving an instruction from program memory that tells the processor what to do. In the decode program step, the instruction is organized into parts that may be significant to different parts or portions of a CPU. The instruction that was fetched and decoded is carried out in the execute program step. An example of an executing program step could be a mathematical computation that instructs a CPU to add input numbers and outputs the final answer. In the final program step, writeback, the processor sends the computed results of the executed steps into some form of memory. This example describes a linear process, the efficiency of which is affected by the time required to complete the four steps before the processor can consider a set of new steps. This bottleneck has been improved over time with the introduction of parallel computing technology where

multiple instructions can be executed simultaneously by sending instructions through different paths in a CPU or by different processors.

Processors may be classified as read only memory or random access memory. A read only memory processor (ROM) is programmed with a permanent collection of routines that will remain in memory if electrical power is lost or turned off. For example, ROM chips send initial startup routines to a computer when it is turned on and retain them when the computer is turned off. Another example of ROM is the common handheld calculator.

Random access memory (RAM) can read and write simultaneously and is the technology that is routinely used to process data routinely. RAM and parallel processing offered significant potential performance improvements that could be realized with the development of the ability for multiple segments of a program to access a computer's memory simultaneously. This advance is known as shared memory. In terms of computer hardware, shared memory typically refers to a large block of RAM that can be concurrently accessed by several CPUs. All CPUs in this type of system share a single view of the data and communications between processors may be fast. While shared memory systems are theoretically the fastest form, some processors may cache, or temporarily store, data, essentially saving unprocessed programming instructions until a processor is free to execute them.

Stonebraker (2007d) considers three categories of database management systems by architecture, one of which is shared memory. Shared memory, the earliest form, was dominant in the 1970s and 1980s. All early relational database management systems were designed for shared memory, including Oracle, Dbase2, Sybase, and MySQL. While fast,

shared memory is the least scalable of the three architectures. The memory becomes a restriction in performance as the number of processors and disks connected to the memory increase.

In the 1990s, a variation of shared memory database system architecture known as shared disk was implemented. In a shared disk system, a collection of processors share a common main memory similar to the original shared memory system. However, a shared disk system uses expanded storage such as storage area networks or multiple hard drives. In a shared disk system, multiple processors have direct access to multiple drives from multiple computers. All computers are able to access all data on the system of multiple hard drives. The disadvantage of shared disk systems is that a single failure of disk hardware can result in data loss. According to DeWitt and Gray (1992), the shared disk design was not considered successful for database applications.

As a result of interviews and discussions with 50 Chief Information Officers and data warehouse administrators, Stonebraker (2007c) reported that, as shared memory and shared disk data warehouse sizes increase, the complexity of query processes that locate data within the system increase exponentially. Non-automated ad hoc queries are used to locate information in data warehouses to support research studies that are more complex than reporting summarized data. Such studies are forensic in nature and are conducted to provide information that may help answer questions prefaced with the words why or how. According to Stonebraker (2007c), the complexity of database architecture and operations reached the point at which many database administrators were refusing to accommodate ad hoc query requests. From a research perspective, the inability to conduct an ad hoc query was a major issue. Arguably, data warehouse systems that are unable to

accommodate essential ad hoc research may constitute database system failure, in that the stated purpose of state longitudinal database systems is to support research on student education and employment outcomes.

Stonebraker (2007a) reported that synchronizing complex systems of shared data is very costly. Although data warehouse procurement may typically cost more than $1 million, system deployment involves integration of data systems connected to the data warehouse, tuning a system for performance, and maintaining a system that is extremely complex. Although integration and tuning are essential, ongoing maintenance of database systems is also essential. Maintenance is considered to be one of the major causes of data warehouse system failures from inability to meet operational needs or the inability to adapt to complex changing circumstances, or the prohibitive cost of operation and maintenance (Sen & Sinha, 2005).

The single point of failure in a shared disk system can be avoided with the third type of database architecture known as shared nothing. A shared nothing system is alternatively known as massively parallel processing (MPP). Each storage section of a shared nothing system communicates with other storage sections for the purpose of replication. If a single disk fails, one of the multiple copies generated in real time are able to reconstruct or replace the failed disk automatically preventing data loss. The MPP design uses a single global file system. Examples of this design are found in products such as Hadoop, Isilon, and IBRIX Fusion. While MPP systems were introduced in the 1980s and all database management systems designed since 1997 use MPP architecture, most of the data warehouse systems in use around the world continue to run on 30 year old technology related to shared memory and shared disk (Stonebraker, 2007a). Oracle,

an enterprise database with an extensive following does not sell a shared-nothing software platform (Hellerstein & Stonebraker, 2005a). Investing large amounts of one-time grant or state appropriated funding for obsolete technology introduces the possibility of performance issues that ultimately must be addressed with total system replacement with little recovery of expended funds.

Limitations of Business Data Warehouse Architecture and Designs

Stonebraker (2007) reported that the then current offerings of the high end data warehouse vendors were "...hard to install, hard to tune, hard to learn, and just generally hard to use. If these products don't get much easier to use then data administration costs will go to 100% sooner or later" (¶ 5). The situation that Stonebraker describes further constrains the ability to use data stored in data warehouse systems as the basis for essential student education and employment outcomes research and analysis.

Stonebraker's (2007) assessment is consistent with earlier and historical criticisms of educational data warehouse systems reported by the U.S. Department of Education (2010) in a longitudinal study. In an interview with a Chief Executive Officer of a project risk management consulting company, Betts (2003) reported causes of IT project failure related to planning and management. In the rush to qualify and allocate billions of dollars in federal grant funding for SLDS, some recipient states' IT leadership may be overlooking risks known by leaders in the IT community and/or may be unaware of the U.S. government caution that the federal funding should be considered a one-time investment (White House, 2009). Making such an investment when needs, capabilities, limitations, and risks are fully understood would seem to be a more prudent alternative to allocating non-renewable grant funds to a high-risk project design.

According to Smith (2004), business leaders may view failure as an unfortunate but necessary part of achieving competitive goals and are willing to risk a series of failures in order to reap the benefits of an innovative business or service that gives a company a competitive advantage. In contrast, "In education, unfortunately, failure is failure both inside and outside the classroom" (Smith, 2004, p. 96).

The history of IT development for education data management systems has notable examples of failure as well. For example, the Idaho state legislature authorized the development of the Idaho Student Information Management System (ISIMS), essentially a statewide longitudinal data system for Idaho. The project was jointly funded with public monies and a foundation grant. The $35 million program design was initiated in 2001 and was terminated as nonperforming in December 2004 when cost overruns and projected development costs reached $182 million. By the time Idaho abandoned the ISIMS project, total expenditures in federal, state, and private grant funding exceeded $24 million.

In his letter to the Joint Legislative Oversight Committee of the Idaho legislature, Mohan, the Director of the Idaho Legislature Office of Performance Evaluations, provided an overview of the lessons learned in the abandonment of the ISIMS project. Mohan (2006) reported:

> Technology projects should clearly define the roles and responsibilities of all stakeholders and consider end users' views, needs, and resources at each stage. In addition, technology projects should maintain a realistic scope, supported by realistic expectations of technology and an updated project plan. The ISIMS project did not adequately address these key issues (p. 5).

The 91 page report documented significant, but unintentional, lapses in the areas of: proper planning, realistic expectations, project milestones, measurable deliverables, competent staff, experienced project management, contract negotiation, compatibility across multiple software packages, centralized systems, small scale testing, management expertise, and issues that could be experienced by other states in the current rush to implement state longitudinal data system (SLDS) projects with grant funding.

State Longitudinal Data System Design Requirements

The Data Quality Campaign (DQC) is a national collaboration that encourages and supports state policymakers to use education data to improve student achievement. In 2007, the DQC surveyed states regarding the development and implementation of SLDS. The surveys indicated progress in the ability to share student records between P-12 and postsecondary systems. However, 34 states reported barriers to aligning systems within their states. The reported barriers included technical issues such as data and data system incompatibilities. Additionally, political barriers were reported that included the lack of cooperation and coordination among staff and stakeholders (DQC, 2009).

The technical and political barriers associated with designing and implementing data systems for government and educational research corresponds to similar issues experienced in business regarding the evolution of information technology systems. Early state database systems were designed to manage personnel and financial records as event-driven data. The systems that managed these data were mainframe-level computers with proprietary applications. Compatibility issues prevented communication between systems (Education Commission of the States, 1998). Over time, demands on these early systems began to exceed system capabilities.

Compliance and accountability reporting are relatively new requirements for education data systems as is the need to store and access student assessment data. However, in general, efforts to build effective compliance and accountability systems would appear to be ineffective and disjointed. According to Carey and Aldeman (2008), "...there's not a single state with a truly comprehensive, effective accountability system" (p. 3) that collects, analyzes, and reports actionable data that can lead to meaningful decisions. They also indicate that "... most states simply gather accountability information and make it available without any clear plan for making it meaningful. Unsurprisingly, it often means far less than it should" (p. 2).

In his testimony to the National Commission on Accountability in Higher Education, Burke (2004) stated that accountability reports often "...appear a grab bag of available indicators with no sense of state priorities or public agenda" (p. 2). According to the National Commission on Accountability in Higher Education (2005):

> ...more accountability of the kinds generally practiced will not help improve performance. Our current system of accountability can best be described as cumbersome, over-designed, confusing, and inefficient. It fails to answer key questions, it overburdens policymakers with excessive, misleading data, and it overburdens institutions by requiring them to report it (p. 6).
>
> A better system of accountability will rely on pride, rather than fear, aspirations rather than minimum standards as its organizing principles. It will not be an instrument for diverting, or shifting blame. It will be collaborative, because responsibility is shared. It will be rigorous, because we can't afford to have low aspirations or soft standards.

> A better system of accountability will be serious about improving performance, while respecting legitimate boundaries between federal, state, and institutional roles, and between policy and educational administration (p. 7).
>
> A better system of accountability will put more emphasis on successful student learning and high quality research (p. 7).

Data necessary to respond to research requirements have evolved into the need to separate and maintain data in multiple data systems. Relating data across multiple data systems and the need for more complex calculations using data have introduced new database design issues. For example, while some personnel and financial data may be updated daily, most of the research data in education and workforce occur on a calendar-driven schedule such as semester, fiscal quarter, fiscal year, or academic year. Descriptive statistics consist of information summarized from event-driven data including totals, maximum, minimum, range, and average (Luan & Willett, 2001). Data systems designed for data storage may support data summaries needed for descriptive statistics; however, data summaries provided by typical data warehouse applications may be insufficient for higher level research. IT staff who operate storage and retrieval systems may lack the training and experience necessary to address fulfillment of increasingly complex research needs that may involve logic, linear algebra, and perhaps calculus in certain applications.

Increasingly, educational research using multiple databases has created demands for statistics across state agency systems that cannot be accommodated with summary data. Luan (2002b), Chief Planning and Research Officer at Cabrillo College in California, presented research at several conferences including the 2002 Annual Forum

for the Association of Institutional Research that demonstrated advanced statistical methods using multiple databases. The clustering analysis example presented by Luan was a technique to evaluate and predict the likelihood for a variety of student outcomes, such as transferability, persistence, retention, and success in classes. At the conference, Luan demonstrated how data mining could predict the individual probability of re-enrollment in the next semester for each student enrolled at that time in a community college in the Silicon Valley. "Compared to traditional analytical studies that are often hindsight and aggregate in nature, data mining is forward looking and is oriented to individual students" (Luan, 2002b, Abstract). Interestingly, data mining methods were initially developed by scholars in higher education to extract useful information and relationships from immense quantities of data. However, business adopted data mining for business intelligence research and currently is the major user of data mining methods.

As emphasis on formal standards and accountability has increased in education, classroom teachers, administrators, and other practitioners are increasingly being asked to use large amounts of data beyond accountability reporting to inform instructional practice, assessment, and evaluation. Terms that describe the new concepts include data-driven decision-making and data for continuous improvement in education (Mandinach, Honey, & Light, 2006). The stated purpose of SLDS is to provide research that supports informed decisions on improving the quality of education and the career potential of students, a shift in emphasis and training is necessary to support the shift from data storage and retrieval, managed by IT staff, to advanced educational research systems for data-driven decision-making such as data mining that are managed by educational research professionals (Luan, 2002a; U.S. Department of Education, Office of Planning,

Evaluation, and Policy Development, 2010). Luan asserts that traditional IT skill sets cannot accommodate advanced research techniques such as data mining. Khabaza (2009) reported that: "Data mining uses advanced technology, and its workings, particularly those of modeling techniques, are unlikely to be understood by the wider IT community" (p. 2). Additionally, data mining projects would not be successful if the investigator is not a domain expert intimately acquainted with the data (Luan, 2006).

Marzano (2003), Senior Scholar at Mid-continent Research for Education and Learning, suggests that proponents of SQL-based data warehouses believe such systems are good at analyzing standardized test data and that standardized test data constitutes effective measures of education achievement upon which decisions regarding educational improvement can be undertaken. Marzano adamantly disagrees with using standardized test results as the basis for data-driven decisions in education and cites studies indicating that using standardized tests can produce false conclusions regarding the effectiveness of schools. Marzano considered state tests a better measure than off-the-shelf standardized tests, but he considered any test that is not related to curriculum goals and objectives to be an invalid indicator of educational effectiveness. On balance, Marzano considered standardized tests to have a place in the "landscape of K-12 education, but schools should not use them as the primary indicator of student learning" (p. 57). Another reason, according to Marzano, is that standardized tests offer little in terms of a plan for interpreting and using the data derived from standardized tests. Henry (2007) agrees with Marzano in that standardized testing "...reduces successful teaching to a single, narrow measure on a multiple-choice instrument" (p. 40). Henry feels that, sooner or later, standardized testing causes teachers to teach to tests in order to survive. More accurate

measures of school effectiveness would involve the study of instructional strategies, classroom management, curriculum design, home atmosphere, learned intelligence in background knowledge, and other factors (Marzano) that may be beyond the reach of business level data warehouse systems. On the other hand, a data system that can effectively analyze data using data mining techniques can accommodate the school effectiveness measures that Marzano believes are necessary. Interestingly, while there is continuing controversy regarding the use of standardized testing as a means of evaluation the value of PK-12 education, Banta (2007) reported a variety of initiatives to extend the use of standardized testing into higher education. According to Banta, the U.S. Department of education in 2007 was promoting a partnership with the National Association of State Universities and Land Grant Colleges and the American Association of State Colleges and University to develop student-learning assessments that would support interstate comparisons of higher education.

Summary

Chapter IV considered issues related to business information technology, including project performance factors and history, the effects of planned obsolescence in business IT hardware and software, the evolution of system design, and limitations of designs in currently used business data warehouse systems. Also discussed were design requirements of state longitudinal data warehouse systems. Chapter V considers design alternatives for data systems optimized for longitudinal research in contrast to business IT systems optimized for data storage, retrieval, and business transactions.

CHAPTER V

LONGITUDINAL RESEARCH SYSTEMS FOR EDUCATIONAL AND EMPLOYMENT OUTCOMES

The emerging need for data-driven decision making beyond the storage and retrieval capabilities of state longitudinal data systems (SLDS) is a foreign concept to many Chief Information Officers and other information technology (IT) staff who may be inadequately prepared for a transition to a different data management paradigm (Data Management Association, 2009). According to Palaich, Good, and van der Ploeg (2004, p. 7), "It is unlikely that any SEA [state education agency] system can fully provide for the data needs of all participants in the education enterprise—and even the most ambitious SEA [state education agency] technical staff should accept this." Until very recently, data management has been largely ignored as a formal discipline in most IT departments. This lack of understanding has resulted in a situation in which the biggest problems in most major IT projects revolve around data integration and information management. Most senior IT managers were not trained in data utilization and analysis because their career focus was generally business data management activities. While senior IT managers should be active participants in research-based data use planning, Palaich, Good, and van der Ploeg cautioned that IT professionals should not lead planning processes.

The shift from data storage and retrieval to advanced statistical methods that require extensive domain knowledge and data mining techniques may create a conflict of

purpose situation that could be resolved through separate leadership of these two functions. In other words, separate leadership within a collaborative environment may be more effective by splitting activities related to data storage and retrieval from activities related to research design, data analysis, visualization, and reporting (Data Management Association, 2009; Mosley, 2008). According to the Data Management Association (2009), data stewards "represent the collective interests of data producers and information consumers" (p. 5) while IT staff serve as data "curators and technical custodians" (p. 5). The analysis and visualization activities in the overall process of converting raw data into actionable knowledge requires extensive domain knowledge that cannot realistically be performed within an automated system without extensive human involvement in preparation (Brachman & Anand, 1996; Cios, Pedrycz, Swiniarski, & Kurgan, 2007; Frawley, Piatetsky-Shapiro, & Matheus, 1992; Romero & Ventura, 2007; Serban & Luan, 2002; Tiwana, 2002). Research design, advanced data analysis methods, visualization, and reporting beyond spreadsheets of tabular data require skills not incorporated into bachelor's level IT education programs or bachelor's level educational programs. Given the distinction in skills sets, it may be more appropriate to consider data storage and retrieval as supporting functions, in a role limited to storing, managing, and providing the high quality data needed by researchers for analysis, visualization, and reporting processes.

Data storage and retrieval processes constitute approximately 20.0% of the work, resources, and time involved in converting raw data into actionable knowledge for decision-making purposes. Analysis, visualization, and reporting account for approximately 80% of the work, resources, and time necessary to complete the

conversion of raw data to actionable knowledge for decision-making. However, most organizations primarily invest in multiple data storage and retrieval systems that severely constrain research (U.S. Department of Education, Office of Planning, Evaluation, & Policy Development, 2010). In extreme cases, 100% of the funding for storage, retrieval, and research related processes are allocated toward data storage and retrieval, essentially the first 20% of the knowledge discovery process necessary for professional research activities that support data-driven decision-making (Cios, Pedrycz, Swiniarski, & Kurgan, 2007). A common trap is obsession with information technology, hardware, and software that some vendors may accommodate with recommendations to adapt new requirements to commercial off-the-shelf (COTS) technology that is renamed to suggest a new marketing initiative more closely aligned to potentially new organizational requirements (Tiwana, 2002). With the primary focus on the technology acquisition, the consequences would appear to be the loss of focus on the research priorities when researchers are not included in planning processes (Palaich, Good, and van der Ploeg, 2004).

As the needs of data-driven decision-making increasingly demand high level and complex statistical analysis of large and multiple databases, a need exists for trained researchers with developed competencies in database theory and design related to research that differs from competencies and skill sets of IT staff (Luan & Willett, 2001). Mandinach, Honey, and Light (2006) emphasize the need to prepare classroom teachers, administrators, and practitioners for future career roles that involve research responsibilities that address the use of database methods for questions such as why groups of students may be experiencing difficulty while other students in the same groups

are exceeding expectations, or how academic improvement plans can be developed that are targeted, responsive, and flexible.

Successful data initiatives are replicable and scalable when they support classroom teachers, administrators, and practitioners at all levels. Wayman, Midgley, and Stringfield (2006) suggested the use of collaborative data teams of data users, including P-20 teachers who study data for classroom improvement application. While teachers may be critical of accountability initiatives, they may embrace the use of data when policies are thoughtfully implemented, respond to the learning needs of students, and are considered useful in improving teaching practices. Educational data management systems designed and implemented by IT staff often ignore the perspective of teachers and other staff who have expressed frustration that data is inaccessible, and when available, the data are unusable for teacher analysis or are formatted in ways that require further education in research and analysis.

According to Wayman, Midgley, and Stringfield (2006), "Twenty-first century school leadership models will undoubtedly be heavy on the use of data to inform decisions. Consequently it is incumbent upon school leaders to identify structures and methods that support the use of student data and involve teachers and other staff" (p. 3).

Preparation of classroom teachers, administrators, practitioners, and other staff for increasingly responsible research roles could begin in pre-service preparation. Colleges of education could integrate using database technology as a tool in teaching and learning. More advanced database coursework could be incorporated into post-graduate programs to address research skills. If implemented, such changes could promote a positive cultural shift in educational organizations responsible for converting information into actionable

knowledge for data-driven decisions where the "ultimate goal is to help institutions to help their students" (Lederman, 2010, May, ¶ 10). In other words, a cultural shift could prepare and empower classroom teachers, administrators, and practitioners who have essential domain knowledge to assume increasingly proactive and responsible roles in using educational data for the improvement of student learning and outcomes at all levels (Wayman, 2005).

State Longitudinal Data System Project Funding

Funding for large-scale SLDS projects may include public, as well as grant funds, with accountability focused on spending according to plan but may lack accountability in demonstrating system functionality with regard to design goals (Idaho Legislature: Office of Performance Evaluations, 2006). A synthesis of literature from many sources indicates that more IT projects fail than are successful (Bodamer, 2010; Charett, 2006, Kanaracus, 2010; Mazon, Pardillo, & Trujillo, 2007; Port & Chen 2004; Tiwana, 2002). The following summarizes statistically unsuccessful factors in project design:

1. Data warehouse design has traditionally been based on underlying operational data sources while overlooking the information needs of decision makers, a scenario that promotes failure in delivering expectations in support of decision making processes. More specifically, failures result from poor communication between data warehouse developers and decision makers (Charette, 2006; Data Management Association, 2009; Goldstein, 2005; Mazon, Pardillo, & Trujillo, 2007; Tiwana, 2002).

2. Development of budgets and spending plans are frequently rushed in order to qualify for funding sources sometimes driven by tight grant submission deadlines (Goldstein, 2005; Strauss, 2010).
3. Budgets are generally aligned with funding expectations, and systems are designed with the help of consultants and vendors toward consuming pre-determined funding limits (Basken, 2010; Tiwana, 2002).
4. Systems generally consist of multiple layers of one-size-fits all COTS hardware and software that require the need for subsequent layers to compensate for constraints within previous layers, a process that introduces additional issues of sustainability, obsolescence, and performance (Alves & Finkelstein, 2002; Basken, 2010; Khabaza, 2009; Tiwana, 2002).
5. SQL software is commonly used in off-the-shelf data warehouse designs that incorporate hard drives that store data in a way that is classified as persistent, in that the data will remain unchanged until revised. In a study of SQL system limitations with regard to different types of queries, Law, Wang, and Zaniolo (2004) reported, "Another well-known problem area for SQL is data mining... It is clear that SQL will be at least as ineffective at mining data streams as it is in mining persistent data" (p.1).

Once the system design and components are established, the typical planning process shifts to designing operational capabilities and policies that accommodate the anticipated constraints of the implemented technology. In other words, capabilities are adapted to technology rather than determining the best technology and best practices to

support necessary operational capabilities (Basken, 2010; Tiwana, 2002; Wayman, Stringfield, & Yokimoski, 2004). While every project is constrained by time, budget, and outcomes, most project plans must be adjusted as a result of unforeseen contingencies that affect one or more of the three constraints. An emerging view of IT project management holds that the project outcomes should be adjusted in order to adhere to the original cost and deadline specifications (Bodamer, 2010). In other words, the full amount of funding is fixed while the implementation of functional capabilities could be decreased to accommodate existing funding limits.

According to Laird (2008), some states may be able to avoid investment in expensive storage, retrieval, and processing systems for SLDS by making minor adjustments to functioning systems. Such a strategy could reduce the risk of project failure with less investment. States reporting experiences with large-scale data systems indicate the need for a research-based approach that could remove compatibility barriers among state agencies and across state governments. A "...data warehouse is not needed to share records... states can share records now by making small adjustments" (Laird, 2008; p. 5).

Need for a Model Longitudinal Research System for Educational and Employment Outcomes

An inexpensive, small-scale model system could be used to develop and test intrastate capabilities across agencies. When sufficiently tested and proven, the intrastate system could be connected to another state's intrastate system to study and resolve interstate compatibility issues. Such a model system could be used in training educational

researchers in colleges of education while at the same time testing usability. This model approach could serve as a catalyst for positive change in database culture from the constraints of COTS data systems optimized for data storage and retrieval toward a research culture that promotes using database technology as a tool at the practitioner level. The need for a culture that respects data as an active part in school improvement processes was addressed in Mason's (2002) two-year study of data-driven decision-making at the classroom level in six Milwaukee public schools. Mason found the need for a culture of classroom teachers, administrators, practitioners, and other staff willing to take charge of data proactively to improve school and educational processes.

A scalable model state level research data system could accommodate Mason's (2002) view of an integrated educational data culture, as well as the development of a functional national research data network if it were designed as a research system instead of a data storage and retrieval system build with COTS software. Dependence on COTS systems creates a situation where designers and users must adapt to technology rather than developing technology to satisfy design goals (Wayman, Stringield, & Yokimoski, 2004). In other words, enterprise databases that use pre-packaged applications from multiple vendors may never provide a centralized solution or have a simple or comprehensive design (Wayman, 2005). In contrast to the poor statistical performance of IT projects planned and executed with information technology design-bid-procure methods a more research-based approach could be used to identify and engineer the technology needed to support research goals rather than adapting research goals to the capabilities of off-the-shelf hardware and software.

Best Practices for Longitudinal Research
Data System Engineering and Design

In a 1995 study of 104 public sector design-build projects, Molenar, Songer, and Barash (1999) reported that 40% were delivered on budget or under budget. Fifty-five percent of the design-build projects studied were completed on schedule or under schedule, while 99 of the 104 projects exceeded design performance expectations. Ninety-two percent of the 104 projects exceeded owner satisfaction expectations. None of the 104 projects failed to conform to specifications. From 1995 to 2004, 1.3% of litigation claims filed against contractors were filed against design-build providers (Design-Build Institute of America, 2010). One of the major differences between design-bid-procure and design-build is that the design-build provider shares more financial risk than a design consultant or vendor who departs the project before work starts. In other words, more financial risk could be a motivator for collaborative success.

Another advantage that could result from implementing a research-based design process in the development of a model longitudinal research data system is the adoption of the formal engineering design methodology. Engineering is the process that creates physical representations from abstract ideas. What distinguishes engineers from practitioners in other disciplines is that engineers apply creative energies toward meeting human needs through design (Khandani, 2005). The engineering design process consists of five phases: Defining problems, gathering pertinent information, studying multiple alternative solutions, analyzing each possible solution to select the best in terms of cost and performance, and then testing the solution before proceeding with full-scale implementation (Khandani, 2005).

Information technology projects planned with traditional IT design-bid-procure project management may have a high statistical risk of failure and may not accommodate educational goals for research, connectivity, usability, accessibility by researchers, scalability, infrastructure support, and resistance to hardware and software obsolescence. Formal engineering protocols and design-build methodology would seem to be a better choice in optimizing SLDS designs for simplicity and function rather than IT designs that stress maximum complexity and support staff (Basken, 2010).

This dissertation has described the on-going history of failures in IT projects suggests that much progress could be made by adopting proven engineering design protocols that address the needs of all levels of users could be more effective than roundtables of IT vendors and consultants determining, with limited study, which combinations of COTS software and hardware would adequately consume predetermined funding goals (Basken, 2010).

Relating the Need for a Model Research Data System to Failures in Business Data Warehouse Systems

Over a period of five years, the U.S. government has authorized almost $5 billion in grants to encourage states to develop and implement SLDS for education improvement, accountability, and research (Mullin & Lebesch, 2010). The intent of the funding appears to be creating an interconnected system of state databases that would accommodate the needs of the U.S. government for a federal data system (Lederman, 2009).

With regard to the complex business data warehousing systems described in the literature that are being considered for adoption in educational SLDS, most fail (Charette,

2006). Further, most education data warehousing systems "were so complex and poorly aligned that their use by school staff was not feasible" (U.S. Department of Education, Office of Planning, Evaluation, and Policy Development, 2010; p. 2).

Specific reasons for the failure of business and education data warehouse systems have been well documented (Charette, 2006; Nemati, Steiger, Iyer, & Herschel, 2002; Smith, 2004; Tiwana, 2002; Weinberger, 2004); however, these focused reports may not completely describe the significance of the consequences of failure. Perhaps various interventions are being implemented to enable partial use of unsatisfactory systems in order to partially protect investments as an alternative to admitting failure. Perhaps many of the systemic problems have been treated symptomatically rather than returning to the drawing board wiser from unpleasant experiences and lost investment. The literature points to the need for a success oriented, research-based engineering approach to building data warehouse systems that will fulfill all needs of education and employment research, as well as the intentions of the governments that are funding the movement to connect state data systems.

While database management systems developed since 1997 use an advanced architecture, most of the database systems currently in use are running on 30-year old technology that Stonebraker (2007d) considers obsolete. Vendor products considered by Stonebraker to be associated with obsolescent data warehouse systems include Oracle, Dbase2, Sybase, and MySQL.

Although integration, tuning, and maintenance are essential, database maintenance is considered to be one of the major causes of data warehouse system failures. Maintenance failures may occur after system implementation but could render

systems inoperable (Sen & Sinha, 2005). A recent U.S. Department of Education, Office of Planning, Evaluation, and Policy Development (2010) study reported a need for education data system models that connect student data to instructional practice. One model for a SLDS, considered by some to support data-driven decision making, is the adaptation of a business data warehouse (Wayman, 2005) connected to multiple state agency data marts in a system developed and managed by IT staff. According to the literature, the statistical probability of success of Oracle, Dbase2, Sybase, and MySQL systems in the U.S. is approximately 17.0% with an overall projected failure rate of 83.3%. Charette (2006) estimates the annual cost of these failed data warehouse systems to be $78 billion.

Summary

Chapter V discussed the complexity of issues related to adapting statistically vulnerable business data warehouse technology to education and employment research that pose greater challenges with correspondingly greater risks and consequences of failure. Chapter V also discussed the need for a model research system that may be capable of responding to complex research needs with more functionally capable and reliable hardware and software that can be operated and managed by researchers. Chapter VI proposes a longitudinal data research system model designed to respond to many of the challenges discussed in chapters I through V.

CHAPTER VI

SIMPLE EFFECTIVE ACTIONABLE KNOWLEDGE OPTIMIZED FOR RESEARCH MODEL

The purpose of this dissertation is to offer a model for a statewide longitudinal relational research data system as an alternative or interim model that may be more effective and less costly than the state longitudinal data systems currently under construction that use complex combinations of commercial off-the-shelf business hardware and software. The Simple Effective Actionable Knowledge Optimized for Research (SEAKOR) model responds to the need for a model education data system reported by the U.S. Department of Education (2010), as well as the needs and issues reported by experts in data warehousing systems in the responsible use of data for education and employment outcomes.

The U.S. Department of education (2010) found that most education data warehousing systems are "...so complex and poorly aligned that their use by school staff was not feasible" (p. 2). In case study research, Lachat and Smith (2005) confirmed the essential need to integrate or link multiple types of student performance data, demographic data, and data regarding students' educational experiences. However, Lachat and Smith reported that the technology necessary to integrate and manipulate data was lacking in school districts, including school districts where extensive data were

maintained. Lachat and Smith found few high schools that had the information technology (IT) capacity to link student results to instructional programs, pedagogy, and classroom environments. Further, teachers and administrators lacked access to enrollment and performance data, and when access was available, teachers and administrators were not always able to determine the effects of programs and practice on student performance over time.

The U.S. Department of Education (2010) determined that the complexity of existing multiple data storage and retrieval products for state longitudinal data systems (SLDS) applications severely constrain research as indicated by interoperability success rates of less than 40.0%. Given credible reports that data warehousing system complexity is related to statistical failure rates as high as 83.3% (Charette, 2006), the SEAKOR best practices model also addresses Laird's (2008) view that a complex, statistically vulnerable "...data warehouse system is not needed to share records... states can share records now by making small adjustments..." (p. 5).

A frequently overlooked method of designing a data warehouse for research purposes is inverting the data storage and retrieval design process steps so that the design proceeds backwards from research outcomes to data instead of trying to identify a use for data already collected and stored without considering research needs, including the need to collect missing data or making corrections to inaccurate data. Because IT led projects typically start with an inventory of data already collected, the collection of existing data tends to drive the rest of the process. In other words, accomplishing the goals of a research project could be affected by inadequate, incomplete, or missing data.

The proposed SEAKOR model will demonstrate a simplified, cost effective, obsolescent resistant system that is capable of evaluating data quality issues, as well as conducting comprehensive research with disaggregate data. The SEAKOR model system could be scaled up or down for implementation at all levels of American education systems to support strategic planning, accountability, school improvement, classroom learning, and reform. The SEAKOR model is designed to fully comply with the 12 mandatory federal elements that define approved SLDS (America COMPETES Act, 2007), as well as the needs and issues reported by Illinois, Indiana, Iowa, Michigan, Minnesota, Ohio, and Wisconsin who are the seven state members of the Regional Education Laboratory Midwest (McDonald, Andal, Brown, & Schneider, 2007). The model supports formative and summative assessment of teaching and learning, as well as the evaluation of curriculum, programs, administration, staffing, and teacher education.

The SEAKOR model is a structured approach to preparing and linking multiple data sources into de-identified, disaggregate individual unit records. De-identified records are individual student records that exist across all levels of education into employment, human service programs, and workforce training programs that have been processed to remove personally identifiable information such as Social Security Numbers and names. The personally identifiable information is replaced across all individual records with a randomly generated identifier that supports linking records across systems without disclosing the identity of any individual. This is essential to conducting research at the level envisioned for intrastate longitudinal data systems while protecting privacy of individuals under the Family Educational Rights and Privacy Act, the Health Insurance Portability and Accountability Act, and the Confidential Information Protection and

Statistical Efficiency Act. If data collection permits, de-identified individual student records may include enrollments, course level information, completion records, and employment outcomes. The SEAKOR model permits multiple occurrences of an individual's de-identified records across multiple agencies to be viewed during analysis, including viewing within a related data portal, while avoiding duplication in statistical analysis and conflict with state and federal laws regarding privacy information. This feature is one of the specific recommendations discussed in the Fordham Law Study: Privacy of Nation's School Children at Risk (Center on Law & Information Policy, 2009).

The SEAKOR model integrates user role-appropriate training into all phases of the process and includes interactive tools for planning, monitoring, and revising data definitions in support of a backward planning process that considers how best to use these data at each level in the education enterprise. The SEAKOR model is scalable and relatively inexpensive in relation to typically used data warehousing systems adapted from business. The SEAKOR model is designed to use successful, though not widely used, technology that could be used to test all phases of interconnected state research data system implementation while saving state and federal resources to fund a new generation of education systems that may enjoy a higher success rate. The model system could test compliance with state and federal laws regarding privacy, test the system security and data validation, test intrastate agency alignment and interstate system interconnectivity, provide an inexpensive model system for research and development of best practices in SLDS design, and demonstrate advanced education-centered and predictive data mining techniques for informed decisions at all levels of education. Further, to support

sustainability, the SEAKOR model could be used in training educational researchers starting with students at colleges of education. This model approach is a low-risk, high-benefit initiative that could serve as a catalyst for positive change in database culture. The SEAKOR model could reduce reliance on ineffective commercial off-the-shelf data systems optimized for data storage and retrieval toward a research culture that promotes using database technology as a tool. Grant funds could be saved to finance new generation systems using functionally proven technology and design as new generation education research systems become available. The SEAKOR model has been engineered to avoid the known risks of failure in the design, development, implementation, and routine use of longitudinal relational research data systems as an interim alternative to statistically vulnerable, extremely expensive data warehouse designs.

The processes and features outlined above are shown graphically in Figure 2. As a system, the SEAKOR model has been engineered to bridge the domain knowledge of classroom teachers, administrators, practitioners, and other staff with supporting tools, protocols, optional processes, and integrated system training. One software application will be used for multiple purposes in varying configurations on various hardware platforms, depending upon desired system scale.

The SEAKOR model will be described in terms of capabilities and processes that transform raw data into actionable knowledge with less discussion of the complexity of hardware and software technology. In other words, typical database system terminology used to describe SLDS capabilities do not necessarily apply to the SEAKOR model due to a simpler design configuration. One of the specific design goals of the SEAKOR model is for end-users involved in data analysis and research to participate in the

development of a localized set of tools tailored to unique needs and interests. Increasing the degree of user ownership of the data system could empower users with greater confidence to use professional knowledge to solve research problems rather than depending on IT staff whose expertise is generally limited to data storage and retrieval and whose support may be limited by their perception of data quality and data warehouse capabilities.

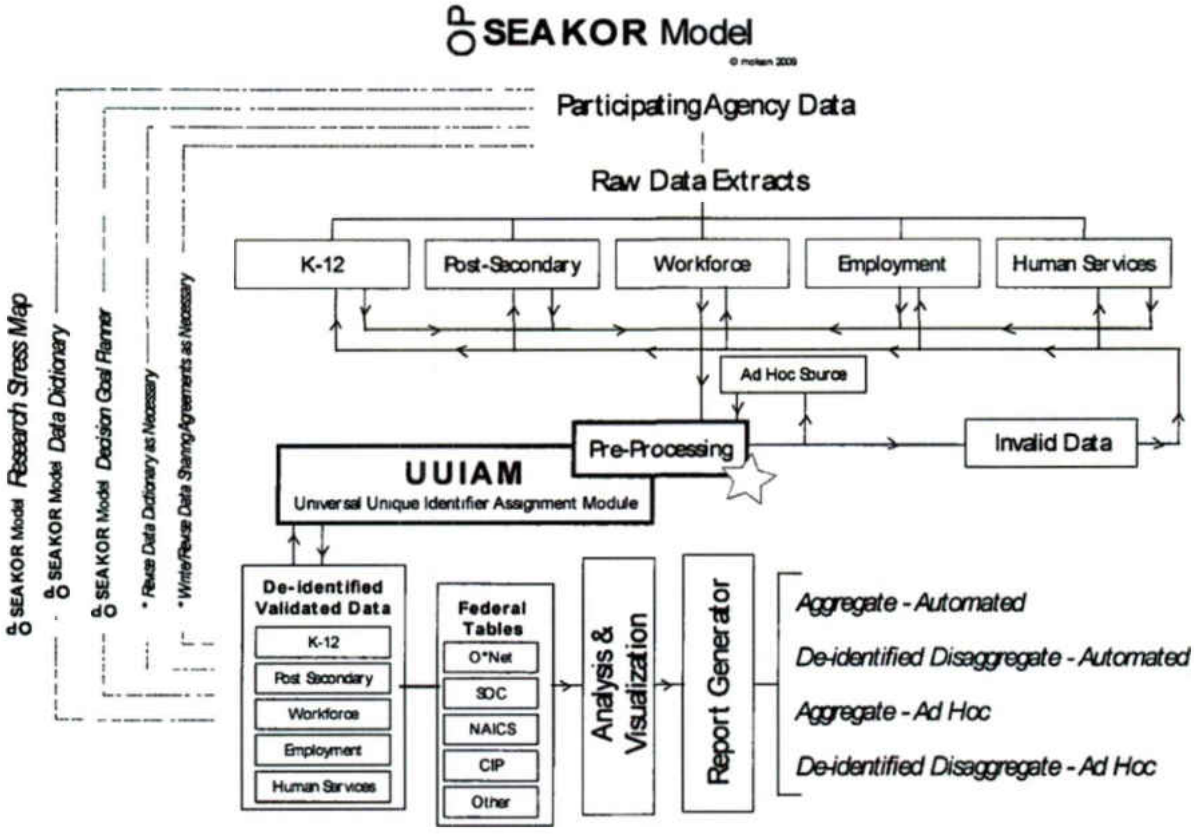

Figure 2. OP SEAKOR model. Data system components and data processing workflow from raw data extracts to completed data analysis report.

Minimum SEAKOR Model System Capabilities

In contrast to SQL data warehouse-related technology that requires dedicated IT staff involvement in research, the SEAKOR model allows non-IT practitioners to focus

on converting raw data into actionable knowledge because of no requirement for IT level computer programming expertise. The SEAKOR model is cross-platform, in that the system will function on Windows and Macintosh computers, including desktops, workstations, and servers as necessary to accommodate capacity or expansion. Using disaggregate raw data extracted from data storage and retrieval systems, all phases of research data storage, retrieval, analysis, visualization and reporting are conducted with one software application.

The SEAKOR model can be scaled to accommodate multiple purpose databases. Each database file can store and process up to 7.75 terabytes of data. Each field within a database table can accommodate 1.85 gigabytes (GB) of data. A database table with 1.85 GB in each field would hold 4,190 fields. In perspective, a single field of 1.85 GB could accommodate 13,000 pages of text that represent 6.2 copies of the 2,074 page American Heritage Dictionary of the English Language. With the advanced server version of the application software, the number of concurrent users is unlimited. The software application has robust security features, is compatible with Macintosh and PC computers, and is capable of two-way live connections with Oracle, Microsoft SQL Server, MySQL data sources, Microsoft Excel, and various delimited data formats. The software supports PHP, HTML, XML, XSLT, and Merge.

A state with a population of 6,000,000 may generate approximately 15 GB of unemployment insurance related data each year that is used by some states for employment outcomes research. At 15 GB of data per year, a single database table in the SEAKOR model system could accommodate approximately 517 years of employment data. As the SEAKOR system is scalable, a second database could be added to extend the

system capacity to 1,034 years. Consistent with industry practices, additional databases could be incorporated to establish a scalable data processing capability. The SEAKOR model system could support preplanned report layouts, as well as more complex ad hoc reports. Container fields can store documents and images that could be used for storing curriculum materials, as well as portfolio materials. The SEAKOR data output can be exported to native database format, pdf, Excel, text, and formats that are compatible with SQL databases including SQL Server, Dbase2, and Oracle.

SEAKOR Model System Security

The SEAKOR model system can be protected with various levels of security options that could restrict access to databases, layout visualizations, calculations, and fields. The security options can incorporate multiple levels of password security, as well as hardware encryption that will prevent access without insertion of a USB slot hardware key. The SEAKOR model system supports automated access logging for monitoring system usage, as well as automated logging and archiving of database revisions. In other words, authorized users could be assigned individual hardware keys. If a key were used inappropriately, a computer monitoring access would deactivate the offending key.

Austere SEAKOR Model System Investment

The minimum, but very capable, SEAKOR model system would consist of one Windows or Macintosh computer with a minimum of a 500 GB hard drive and one software application package, assuming the organization is currently using Adobe Acrobat Professional, MicroSoft Office, and SPSS. A larger hard drive, including external drives, could support more frequent data backups and larger file sizes. The approximate cost of the software application is estimated between $300-500 and the

approximate cost of necessary hardware is estimated between $2500-3000 for a system that permits expansion. Actual cost would depend on configuration needs and government or education discounts.

The following sections address the SEAKOR model design and implementation processes shown in Figure 2, including processes related to identifying data necessary to support research studies and decision outcomes.

SEAKOR Model System Research Stress Map

Many classroom teachers, administrators, practitioners, and other staff, even those with limited experience, may intuitively sense that success on a single standardized test is not necessarily cause for celebration, just as a single set of unsatisfactory test results may not necessarily justify reorganizing a school. In other words, many factors form a dynamic system of force vectors working from different directions that make it difficult to reach conclusions with limited data. In more direct language, "...standardized test scores cannot provide data at the depth and frequency necessary to inform decisions about instructional practice" (Palaich, Good, & van der Ploeg, 2004).

Many of those involved in researching the outcomes of education and employment become increasingly aware of the complexity of the process and the amount of data that must be sifted and sorted across multiple state agencies and databases to find and link the data necessary for analysis. Individuals whose analytic experience is limited to the data in a gradebook or the use of spreadsheets to analyze student grades or standardized test scores will not enjoy the same perspective as those who work with large amounts of data. In other words, gradebook scores and standardized test results constitute a limited cross-sectional look at performance uninformed by academic performance

occurring before or after the cross-sectional snapshot or other educational performance indicators. Before an individual can embark on the process of learning how to use a database for developing comprehensive information that can support instructional or policy decisions, it is necessary to understand the capabilities of databases in the continuum of research methods and school reform.

To assist researchers, particularly new researchers, with understanding the nature of localized education data, the SEAKOR model database design process includes a step that involves the development of a localized research stress map to orient researchers to the sources of available data and potential issues in linking data across data sources. The Research Stress Map is an aid to understanding the need for database technology and the gaps through which information could fall if not captured in the research process. Figure 3 is a sample stress map that would be developed by a SEAKOR implementation team that would include a facilitator and organization members being prepared for SEAKOR implementation. A subtly important goal of the research stress map is reduced stress that may come from knowing that a research data system is not a euphemism for laying blame on teachers. The research stress map can become a graphic realization that teacher led activities in a classroom have a profound effect on students, but teacher led activities are not the single source of profound effects that can mean the difference between success or failure of both the student and teacher. Some voices in education or voices in positions to influence education may have developed the notion that data-driven decision-making is about placing blame for poor results instead of realizing the positive benefits of doing a root cause analysis of a phenomenon in order to introduce a positive intervention without

blame. Teachers are concerned about blame for situations beyond their ability to influence.

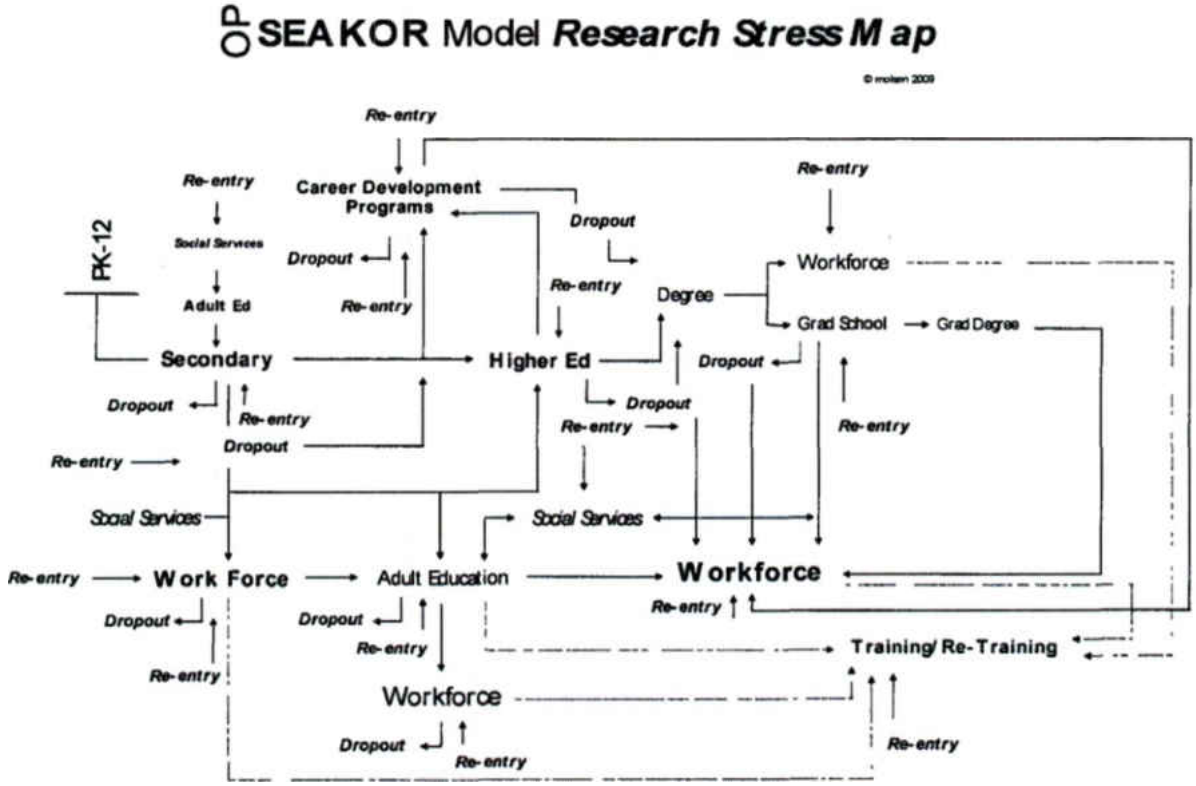

Figure 3. OP SEAKOR model research stress map. A representative research process illustrating the complexity related to following students across state agencies and data systems from secondary education to employment outcomes.

According to McLeod (2005),

> One of the most important things administrators can do to foster data-driven educational practices is to facilitate school climates where it is professionally and emotionally safe to look at student data. Teachers will resist using data if they feel that the information will be used against them for evaluative or punitive purposes (¶ 31).

Carey and Aldeman (2008) indicate that "... accountability, implemented correctly, can be an asset instead of a threat" (p. 4). Several of Deming's (2000) fourteen points for quality agree with McLeod (2005) and Carey and Aldeman (2008): Eliminate quotas, break down barriers within the organization, and "Drive out fear, so that everyone may work effectively" (Deming, 2000, p. 23). Essentially, root cause analysis uses a set of formal procedures to separate symptoms from the underlying cause of a phenomenon. Root cause analysis philosophy ignores the politics of expedient decisions in order to determine the reason for an occurrence so that it may be eliminated. In contrast, organizations, including government entities, often focus on what is most visible, where the deepest pockets of money are, or whatever is politically convenient toward taking remedial actions that may make problems less visible through a patchwork of incomplete or temporary solutions (Okes, 2009).

A hypothetical case provides an example of a data-driven decision-making issue that could be studied with the SEAKOR model. The situation involves a group of students from the same rural area who enrolled as freshmen in the same university following high school graduation. The group represents approximately 65.0% of the students from the same high school. During placement testing, the university discovered that every member of this subgroup of students required reading remediation due to low scores, and the concern is sufficiently significant to notify the school involved. It was discovered that the students had the same group of reading teachers in elementary school, and the blame began to shift toward those teachers. Before passing judgment, thoughtful research using database technology could perhaps discover that, while this group had the same elementary reading teachers, all of the students who required remediation had

significant absenteeism problems unrelated to the teachers' performance. If there were an effective SEAKOR model system, or other failure resistant database research system with the ability to follow individual records for longitudinal research at the time that the subgroup of children were attending elementary school, the outcome could have been different. Data mining techniques could have been used to discover a pattern in the social issues, attendance records, or other matters related to subsequent degradation in reading performance years before enrollment in the university as freshman. In other words, data mining could have possibly identified issues unrelated to the incorrect assessment of poor pedagogy. If identified early and remediated, the need for remedial reading at the post-secondary level could have been avoided.

SEAKOR Model Data Dictionary

The purpose of a data dictionary is to define the properties of the data stored in a database system. Data properties in a data dictionary include data elements, attributes, units of measure, and other information related to standardizing quality and to develop the means to translate data in cases where data does not meet standard definitions, including standardized definitions across multiple databases. While data dictionaries were originally used by IT staff involved in designing data storage and retrieval systems, they are increasingly important in guiding policy makers, data analysts, and researchers in activities such as direct data input and query writing for data extraction. This higher-level use for data dictionaries has created a need for a higher-level structure with expanded information that is necessary to assist end users in understanding the data structures and contextual meaning. If a database dictionary exists and is accurate, it can assist in troubleshooting symptoms of unusual data patterns. For systems that are connected across

regions or states, local data dictionaries would be necessary to align databases and reduce redundancy.

Essentially, during implementation of data storage and retrieval systems, data dictionaries are established to ensure that information is stored in the appropriate location for retrieval and analysis. If a data system stores information on ACT test results that are intended to measure college readiness and there is no procedure to ensure that scores for English, mathematics, science, and reading are stored in separate locations, then analysis of ACT data would lead to invalid results.

Critically important is using a data dictionary to support a bridge between a research question and the data needed for analysis. Incredibly, the development of data dictionaries may be omitted during the implementation of data storage and retrieval systems in the interests of expediency leading to situations in which ten to twenty analysts and programmers could be gathered in a room to identify the location of specific data needed for a statistical study.

Another very important use for the data dictionary, especially in cases where data systems must be aligned for a common purpose, is to provide guidance to translate differences in data across databases to support linking of data for analysis using common definitions. For instance, one agency's legacy system may export dates as numbers while other systems may export dates in various data formats. The SEAKOR model system is designed to remove a barrier to interconnectivity, in that SEAKOR does not enforce arbitrary standards for communication among data systems. Data translator modules may be used to convert data in various formats into a uniform format that permits direct comparison. Experience has shown that enforcing unnecessary and unrealistic standards

across data systems could be cost prohibitive for many organizations because the data formats in use may affect multiple systems within an agency so that changes would introduce cascading effects.

A key component of the SEAKOR model is the process of developing a comprehensive data dictionary that meets the needs of researchers, analysts, and other users with authoritative and accurate field data definitions. Carefully defining data elements used in computations or linking data systems is the first step toward beneficial use of data (Northwest Environmental Data Network, 2006).

SEAKOR Model System Decision Goal Planner

Essentially, the decision goal planner process is a bridge that connects data users to the sources of the data. The SEAKOR model design is intended to engage researchers, analysts, instructional staff, and decision makers in working backward from the decisions that would be made with data to improve the quality of teaching, learning, and employment outcomes. If correctly executed, the process will define the information needed for strategic planning, accountability, and educational process decisions at all levels. The data required to fulfill these needs is described in a comprehensive data dictionary. A data dictionary could assist in the identification of the need for data that do not yet exist or not yet collected. A data dictionary is also helpful in identifying data definitions that do not describe the data being stored. In such cases, new or revised data definitions are created to guide the collection of the necessary data. Further, the process will define the path from a data-driven decision, including classroom decisions, to the data required for analysis. In contrast, some software vendors and consultants tend to assure the inexperienced that all that is needed to obtain research data is simply type in

the question in plain language, and the computer will give a complete data analytical report.

Brooks (1987), Professor of Computer Science at the University of North Carolina, internationally recognized computer scientist, and the recipient of the National Medal of Technology wrote that people for 40 years had been writing about automatic programming, meaning the ability to solve a problem from stating the problem. Brooks (1987) states that "...it is the solution method, not the problem, whose specifications has to be given" (p. 8). The SEAKOR model system accommodates Brooks' analysis, in that the backward decision goal planner, as shown in Figure 4, reveals the solution method that is then mathematically described to the SEAKOR system to analyze data. As the fields needed for analysis are identified, their field definitions in the dictionary should be verified to ensure that the data definition, elements, and attributes are appropriate for the analysis required to satisfy the decision goal of the process. This validation will ensure that the correct data is used to avoid reaching an incorrect analysis. Once the decision goal planner is completed and tested, and assuming there are no changes in the path backward from decision to data, it may be possible to automate the specific research method to produce a repetitive custom report in cases where it may be appropriate to do so. The backward decision goal planner is also recommended for ad hoc reports even though they may be required only once. The reason the backward decision process is recommended is to provide a validated audit of the process used to determine the specifications for the ad hoc study.

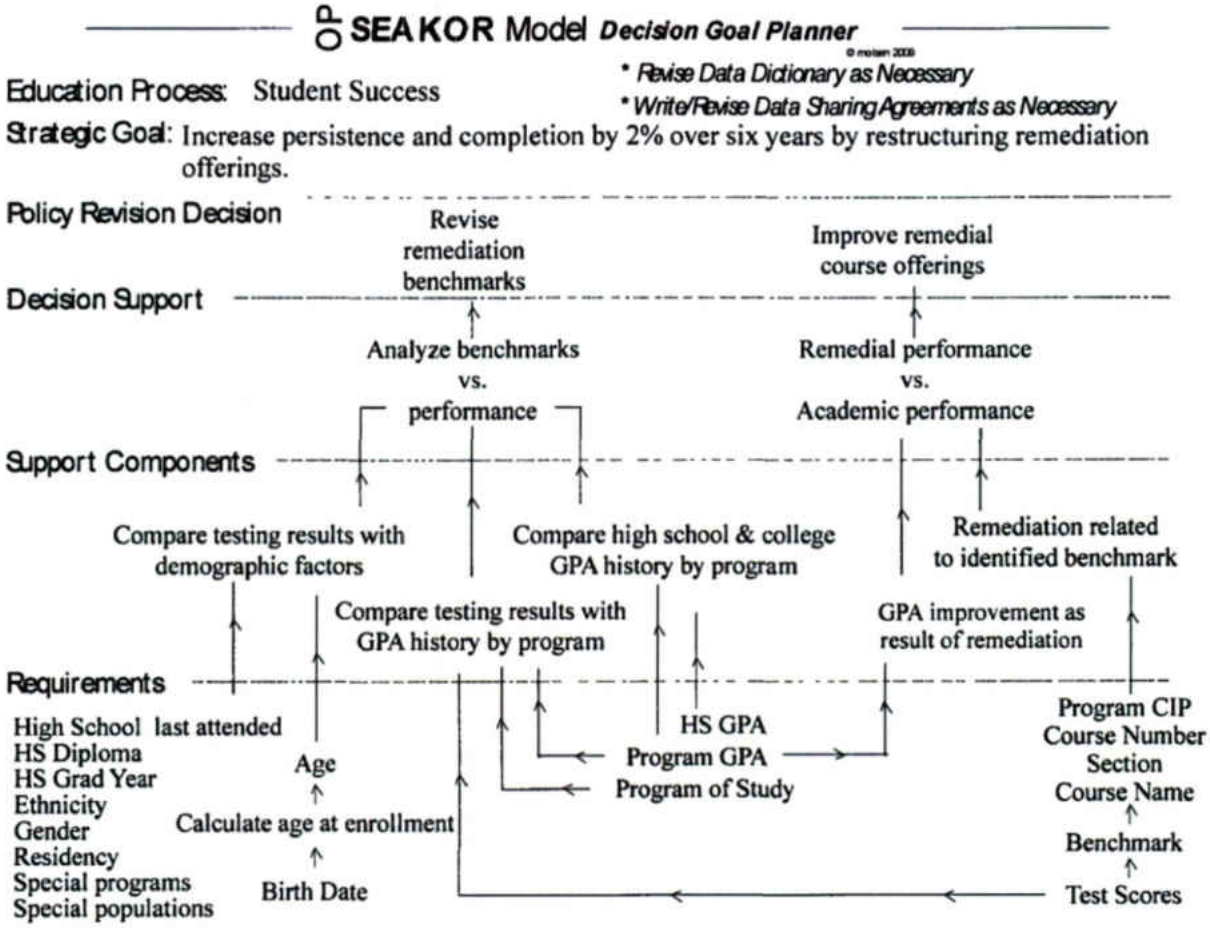

Figure 4. OP SEAKOR model decision goal planner. A representative model research goal planning process that starts with a desired research outcome and proceeds through a process that is completed with identifying the data fields.

When fully implemented as a process, the information contained in the organization's collection of decision goal planners would be entered into an expanded, interactive research dictionary that, in separate tables, provides related views of data fields, support components, decision support information, decisions, and strategic goals or response to high level research questions. A SEAKOR model interactive research dictionary could store specific data sharing agreements related to specific research

projects along with the ability to access the specific data sharing agreements that may apply to a specific research project. In other words, it is possible to access the field definition table and identify the strategic planning requirements that use those data fields. Alternatively, it is possible to select a report from a list of reports. When a report is selected it would then display all of the data fields for that report, as well as provide the means to verify the accuracy of field data definitions, other steps in the decision goal planner, and the status of related data sharing agreements. The interactive tool is designed to avoid entering the same data more than once.

The goal planners are not intended to be rigid templates but rather a conceptual framework for working backward from the outcome of research to the data considered necessary to include in the chosen study. Goal planner methodology would be implemented in the training phases of the SEAKOR model implementation so that teachers, administrators, and policy makers can discuss educational research processes with a commonly accepted taxonomy of terms.

SEAKOR Model System Data Extracts

The SEAKOR model differs from a traditional data warehouse that accepts disaggregate data from data marts for aggregate data reporting. The traditional data warehouse system requires an extraction point between data marts and the data warehouse in order to collect disaggregate data because data warehouses are limited to aggregate data reporting. The SEAKOR system eliminates the traditional data warehouse because disaggregate and aggregate data processing are accomplished in the module that includes analysis, visualization, and report generation. In other words, disaggregate data extracts are collected directly from data marts that contain disaggregate data.

The data extracts are predetermined record data that consist of the data fields that have been defined as necessary to support the research requirements based on an appropriate extract schedule that may be different for different agencies. In general, the raw data extracts are calendar-based exports from agency databases or data marts. An example of a calendar-based extract would include records of postsecondary program completers at the end of an academic year or semester. Another example of a calendar-based extract would be postsecondary enrollments associated with specific semesters. If employment data is based on unemployment insurance data, these data may be extracted on calendar quarter or calendar year basis. Collecting these data generally includes a lag time during which owner agencies perform various validation processes. For calendar-based data, the data extract collection method serves the purpose at minimal expense. In other words, there are no benefits to continuous collection of calendar-based data, and there may be disadvantages to collecting calendar-based data on a continuous basis if various researchers are accessing the same data at different times for different purposes. In such cases, results could be different causing bias or skewed research results.

In their analysis of U.S. government legislation implemented to link education with employment outcomes, Mullin and Lebesch (2010) reported

> ...the lessons learned from workforce legislation are beneficial to the way states think about measuring performance from all segments of education, not just workforce programs....It may be that statewide data exchanges share data between sectors of education and the workforce in a periodic and systematic way (p. 9).

Other aspects of data pre-processing are included in the next paragraph.

SEAKOR Model System Data Pre-processing Protocols

The literature consistently reports concern with providing disaggregate data in ways that could violate federal and state privacy laws if personally identifiable information were included in record data. However, at some level, records cannot be successfully matched across databases or agencies without the use of an identifier that is common across all the records being matched from various sources. In SEAKOR model terminology, records that are matched across databases or agencies using a common identifier are known as related records. Fields from related records can be viewed in a composite record that includes appropriate data from all related records, a feature that provides a catalyst for data analysis, as well as related record validation. An example of this situation could be matching K-12 records with a system identifier to higher education records that may contain a social security number or a different system identifier to employment records that most likely use social security numbers as unique identifiers. Therefore, it is not realistically possible to match K-12 records, higher education records, and employment records without a common identifier that is attached to all such records of a specific individual, a process that most existing data warehouse systems cannot successfully accommodate without additional software. The SEAKOR model system is designed to avoid such issues while maintaining the capability to aggregate data, sub-aggregate data, or use disaggregate data in analysis and reporting without disclosing personally identifiable information.

In the SEAKOR model system, raw data records originating across several agencies are maintained in their own encrypted data tables with all originating identifiable information, such as social security number, system ID, name, birth date, and

birthplace if available. Also within a secure environment, records are automatically validated. Invalid records are identified and owner agencies are automatically informed. Within the pre-processing step, metadata is validated or created as necessary.

SEAKOR Model System Universal Unique Identifier Assignment Module

The purpose of the Universal Unique Identifier Assignment Module (UUIAM) in the SEAKOR model system is to manage the assignment of a unique identifier attached to all records that would distinguish each individual's records from all other individuals' records. For instance, if the historical records of an individual were to include two program completion records, eight enrollment records, and twelve employment records, 22 records within SEAKOR would be identified with the same unique identifier different from all other identifiers. The UUIAM accommodates data mapping requirements, in that each of the 22 records for the hypothetical individual could be mapped to the table where the records are stored using the assigned unique identifier.

The UUIAM is protected with robust security with options that restrict access to the databases, layout visualizations, calculations, or field data. The security options include multiple levels of password security, as well as the capability to use a hardware encryption tool that prevents the UUIAM database from being accessed without proper password, as well as insertion of a coded hardware key into the associated computer's USB slot. This feature supports automated access logging for monitoring the UUIAM usage. In other words, a very restricted set of authorized users could be assigned individual hardware keys for situations that required authorized use of privacy information. If someone attempted to use the hardware key and password inappropriately,

a secure computer assigned to monitor system access would immediately deactivate the offending key.

These procedures may seem to be extreme measures for a database system already protected by limited access. However, these procedures are justified by the dismal security record that exists for educational data systems in the U.S. Therefore, data security of education related data systems has become a trust issue. For the years 2006 to 2010, the Privacy Rights Clearinghouse (2010) reported 372 breaches in educational systems representing 5,624,105 privacy records. The Fordham Center on Law and Information policy (Center on Law & Information Policy, 2009) recently reported a lack of privacy protection in longitudinal databases in the majority of states. For example, 32.0% of the states' data warehouses failed to protect student social security numbers, 22.0% of the states' data systems failed to protect information related to children's pregnancies, and 46.0% of the states failed to protect information related to mental health and jail sentences in children's educational records. The Fordham study (2009) also found that the flow of information from educational agencies failed to comply with the provisions of the Family Educational Rights and Privacy Act. Additionally, the Fordham study found that several out-of-state contractors hired to warehouse states' student data failed to protect the data entrusted to them. A more critical dimension exists with vendor disclosure of privacy data, in that the responsibility of data security lies with the owner of the data. A contractor or vendor that knowingly or accidentally discloses privacy data would escape consequences without established contract provisions. However, the Fordham study found cases of contractors and vendor having custody of state student data without contracts defining responsibilities. In other words, the responsibility of data

a secure computer assigned to monitor system access would immediately deactivate the offending key.

These procedures may seem to be extreme measures for a database system already protected by limited access. However, these procedures are justified by the dismal security record that exists for educational data systems in the U.S. Therefore, data security of education related data systems has become a trust issue. For the years 2006 to 2010, the Privacy Rights Clearinghouse (2010) reported 372 breaches in educational systems representing 5,624,105 privacy records. The Fordham Center on Law and Information policy (Center on Law & Information Policy, 2009) recently reported a lack of privacy protection in longitudinal databases in the majority of states. For example, 32.0% of the states' data warehouses failed to protect student social security numbers, 22.0% of the states' data systems failed to protect information related to children's pregnancies, and 46.0% of the states failed to protect information related to mental health and jail sentences in children's educational records. The Fordham study (2009) also found that the flow of information from educational agencies failed to comply with the provisions of the Family Educational Rights and Privacy Act. Additionally, the Fordham study found that several out-of-state contractors hired to warehouse states' student data failed to protect the data entrusted to them. A more critical dimension exists with vendor disclosure of privacy data, in that the responsibility of data security lies with the owner of the data. A contractor or vendor that knowingly or accidentally discloses privacy data would escape consequences without established contract provisions. However, the Fordham study found cases of contractors and vendor having custody of state student data without contracts defining responsibilities. In other words, the responsibility of data

security stops with the owner agency. Fordham's findings reflect confusing signals and, perhaps, the lack of leadership at levels of the federal and state governments that are responsible for implementing SLDS that are required to comply with existing federal and state laws.

To avoid issues related to unauthorized disclosure of privacy information, the SEAKOR model system is designed to restrict disclosure of privacy information through procedures that create de-identified records used for research, analysis, and reporting. Individual records within the UUIAM are matched using identifiers provided by the owner agencies. Matched records are processed so that the records being imported into the UUIAM are either assigned a unique identifier or are annotated with a previously assigned unique identifier that is subsequently used to match records across databases and agencies without risk of disclosing identity. In other words, records from K-12 education are imported into the de-identified validated data tables after removing all personally identifiable information while retaining the anonymous identifier. Similarly, employment records are imported into the de-identified validated data tables after removing all personally identifiable information while retaining the anonymous identifier. These procedures make it possible to match records with K-12 records without disclosing personally identifiable information. All agency data, upon the completion of pre-processing and UUIAM assignment or validation would be transferred to the de-identified validated data tables for use in statistical processes.

The process is tested to guarantee anonymity, as well as uniqueness of each individual's record. The benefit of this procedure is to collect validated record data in disaggregate form that can be analyzed in disaggregate form, sub-aggregate form, or

aggregate form, based on need for any statistical process, including multiple use for multiple purposes.

SEAKOR Model System Validated De-identified Research Data

The records stored in the validated, de-identified research data tables can be used for analysis and reporting that require aggregate records, sub-aggregate records, or disaggregate records without risk of disclosing personally identifiably information. These data are imported from raw data provided and updated by owner agencies based on event driven calendar schedules appropriate for various data types.

Integration of Federal Data Systems within the SEAKOR Model System Interface

The SEAKOR system model is designed to match records with federal databases related to Classification of Instructional Programs, Standard Occupational Codes, occupational descriptions (O*Net), and North American Industry Classification System codes. The U.S. government provides this information in downloadable tables that work well with calendar-based data extracts. These data are necessary to determine how a program of study may be related to occupations or industries.

SEAKOR Model System Data Processing Capabilities

The SEAKOR model system was designed for classroom teachers, administrators, practitioners, and other staff from varying backgrounds that may or may not include specific computer science or computer programming education. The requirement for SEAKOR user competence is the ability to specify computer instructions using mathematics and logic. The SEAKOR model data system may be visualized as a very flexible array of data in rows and columns without the restrictions of a spreadsheet

application. Tools within the system allow reports to be formatted as desired. Fields are user configured to be text, number, date, calculation, summary, and container. A container field will store documents, images, and other types of files.

In relation to typical data warehousing systems that involve multiple applications, including complex applications such as Oracle or Dbase2, the SEAKOR model system uses one application programmed with a mathematics based and object oriented interface. In other words, any requirement that can be described logically in mathematical notation can be accomplished with the SEAKOR model system. For example, a field can combine dates, numbers, or text from multiple tables linked through relationships using mathematical equations to compare and evaluate information contained in fields or groups of fields in a system of related tables. An example could be selecting a subgroup of individuals who are 25 years old or older who are enrolled for the first time in a postsecondary program of study identified as a high demand occupation according to the U.S. Bureau of Labor Statistics (2010). Another example could be determining whether students who completed a program identified as a high demand occupation are employed in an occupation related to their program of study five years after graduation. Then, in the same step or a subsequent step, the subgroup could be analyzed to determine median income based on the most significant source of income and the number of graduates employed in related occupations working full-time, part-time, or multiple jobs.

Similar techniques could be used to learn about the outcomes of students who did not graduate in terms of patterns that could be studied to determine support mechanisms that may help students persist to degree attainment. Using advanced techniques within the capabilities of the SEAKOR model system, data such as these could be analyzed to

determine the probability of an individual completing a specific program of study and finding employment in a related occupation over periods of time, just as data could be analyzed to predict program completion prospects for students from different backgrounds and educational situations. Other examples could include analyzing student outcomes to determine the effect of demographics, schools, and programs. Student data could be further analyzed to identify characteristics of students who achieve proficiency in relation to students who do not achieve proficiency. Student data could also be studied to evaluate how closely state test results correlate with actual classroom performance.

Stilton (2010) stated curiosity as to why classroom and teacher assignments, course choices, discipline and suspension issues, honor performance, as well as other aspects of student activities could not be centrally stored to provide data for generating trends and other reports necessary for making sound data based decisions. Stilton's concept of data collection and analysis can be accommodated with relational research systems such as SEAKOR that support advanced statistical techniques for large databases. However, without training, these or other research possibilities could be overlooked, especially by those whose backgrounds primarily involve the use of spreadsheets or data printouts provided by IT staff.

The process used in the SEAKOR model system to prepare data for analysis is called Knowledge Discovery in Databases (KDD), a data processing technique pioneered by Fayyad, Piatetsky-Shapiro, and Smyth (1996). The KDD processing technique consists of five steps, the fourth step of which is advanced statistical methods for very large data sets. These methods are known as data mining. The five steps include data selection, data pre-processing, data transformation, data mining, and data interpretation or

evaluation. Together, these five steps define a process that starts with raw data and results in actionable knowledge. The fourth step, transformation, is a key process that allows data defined differently in different data sets to be translated or interpreted to common terminology that related databases can recognize and use in calculations without expense and significant risk of changing field definitions in traditional data warehouses.

The SEAKOR model system uses a structured approach to applying the KDD process in ways that assist users with pre-planning the preparation of data to support strategic goals and decisions. The SEAKOR approach involves working backwards from the goals of research through intermediate steps to defining the data fields necessary to support achieving the research goals outlined in the decision goal planner process, as seen in Figure 3. Essentially, the decision goal planner process is a bridge that connects data users to the purposes of the data.

A barrier to the success of the SEAKOR model or other processes that support analysis of data for educational decisions is the ability of IT staff to extract required disaggregate field data in homogeneous format that has a key field to support matching across data tables. Examples of key fields would include social security number, system identifier, last name, first name, middle initial, birth date, and birthplace if available. In other words, key fields and essential contextual fields, such as term of enrollment of graduation date, distinguish a record or matched records from all other records.

Using an additional data attribute such as Federal Employer Identification Number could be used to distinguish between separate sources of income if an individual were to have two or more sources of income. Using another attribute, such as, postsecondary education institution and/or classification of instructional program code

could result in the ability to map student choices or changes in thinking regarding career education through the ability to distinguish among different programs of study at different times in different institutions. Analyzing such data requires the use of logic that typical data warehouses are unable to accommodate due to the inability to perform computations using Boolean logic. A more complicated but possible analysis could be comparing the employment potential of students' most significant source of income across the most significant programs of study in cases of multiple degrees and multiple sources of income.

SEAKOR Model System Data Analysis Visualization and Reporting Capabilities

McDonald, Andal, Brown, and Schneider (2007) reported significant concerns of the Midwest region states regarding the use of data systems to improve education, processes and outcomes as the:

1. Ability to protect individual privacy rights.
2. Ability to analyze disaggregate data.
3. Ability to extract information for multiple purposes.
4. Ability to eliminate duplicative data collections.
5. Ability to report the same data in multiple forms.
6. Ability to enrich information already analyzed by relating additional information.
7. Ability to develop regional benchmarks.
8. Ability to develop analytic resources, as well as training materials to support expanded data analysis and reporting.

A major design goal for the SEAKOR model was the ability to respond to unique local or regional needs and barriers with the ability to successfully connect systems with different data definitions using data translation or interpretation methods to prevent expensive modification of existing systems. The following describes how the SEAKOR model could respond to the priorities of the Midwest region with the flexibility to make adjustments at any point at any time. In other words, the SEAKOR system is not constrained by technology or limited in the ability to adapt to changing circumstances beyond the necessity for disaggregate raw data with identifiers that support matching records across systems and agencies.

With regard to the ability of the SEAKOR model to protect individual privacy rights, the UUIAM module assigns an identifier to extracted records at the time records are imported into designated research data sets. The personal identifiers are removed and maintained in a highly secure encrypted and monitored environment, access to which requires a specific data sharing agreement, a password, and a hardware key. In other words, without specific authorization, SEAKOR system users would be unable to view designated personally identifiable information because these data would be removed from data imported into the SEAKOR's report generator for analysis.

With regard to the ability to analyze disaggregate data, the validated and de-identified raw data imported into the SEAKOR de-identified validated data research data sets are maintained in disaggregate form, including multiple occurrences and multiple records where they exist. The multiple occurrences and multiple records may be selectively aggregated, sub-aggregated, or used as disaggregate records depending upon

the selected research methodology. An example could be counting individuals who completed a degree program as compared to counting degrees completed.

With regard to the ability to extract information for multiple purposes, the designated research data sets are repositories of single purpose longitudinal data. Examples may include enrollments, course data, and employment. While the data in these repositories are used for multiple purposes, the system is designed so that the source data remains unaltered for future use. Therefore, the SEAKOR model allows unlimited access to these data sets for calculations or relationships in response to multiple purpose research questions without altering the source data.

With regard to the ability to eliminate duplicative data collections, the SEAKOR model system is a tool that can be implemented within a school, district, university, university system, multiple state agencies, or within a state government environment with the capability to respond to local needs, as well as the needs of any SEAKOR system to which the local data system is connected. The SEAKOR system is designed so that the data collected at any level is collected once as the basis for extraction and importing into a SEAKOR system. Assuming effective backup of data stored in SEAKOR's data sets, there is no further duplication, as the data set is the basis for all research.

With regard to the ability to report the same data in multiple forms, the SEAKOR model allows the same validated and de-identified raw data to be accessed for different reports. Report formats may be saved and automated so that multiple reports using the same data can be automatically produced at the same time or on a calendar schedule.

With regard to the ability to enrich information already analyzed, a modified analysis can be accomplished as an ad hoc process added to a saved report or,

alternatively, a decision goal planner can be modified to add additional processing requirements that would result in a saved report format for possible automation.

With regard to the ability to develop regional benchmarks, more than one alternative exists. If multiple states use an interconnected SEAKOR systems, a state could access all of the comparable data that exists across the region for analysis. For example, a state that uses ACT, could access ACT data across the region. Similarly, a state that wanted to compare Iowa Test of Basic Skills data could access the data across the interconnected states for comparison in related tables. If states were not interconnected, such research could still occur by sharing extracted and de-identified data via any means because privacy data has been automatically removed from records before importing into SEAKOR. In other words, the SEAKOR system was designed to avoid security issues related to personally identifiable information.

With regard to the ability to develop analytic resources and training materials, the SEAKOR system is designed to respond effectively without IT resources beyond data storage and retrieval activities because the system consists of one software application. Additionally, the SEAKOR system installation is the result of an organizational collaborative design effort. Proper implementation includes integrated training for all user levels with training materials and facilitators. The goal is to create a culture of organizational self-sufficiency with regard to data analysis, visualization, reporting, data-driven decisions, and strategic planning.

The SEAKOR model system has extensive and flexible data analysis, visualization, and reporting capabilities. The capabilities include defining and exporting data sets for traditional statistical analysis using common statistical applications. The

SEAKOR model system was also designed to efficiently conduct research, analysis, visualization, and reporting of data that involve relationships, including longitudinal relationships in multiple tables. These methods accommodate data tables of very large data sets using KDD techniques designed for projects of this magnitude. The formal structure of the first three steps of the KDD process is necessary to prepare data for step four, data mining techniques, a complex step that requires extensive domain knowledge, as well as the possibility of linear algebra and calculations that could include calculus. An example of such a study could be predicting the number of freshman dropouts based on performance, remedial course work, extracurricular activities, economic status, employment, test scores, high school history, and other socio-economic factors. While the need for mathematics skills could be intimidating for some users, the logic and mathematics interface is significantly less intimidating than the effort involved in mastering multiple software packages with different coding systems with data warehouse techniques that have a low probability of success.

In a state with a population of six million, a SEAKOR model system could be called upon to analyze an array of related tables that define educational and employment outcomes over a period of years. If the longitudinal analysis included five years of data, this could represent the use of raw data sets with a total of approximately 75 million employment records, six hundred thousand post-secondary records, several million miscellaneous state agency records, and federal database tables related to Classification of Instructional Programs, Standard Occupational Codes, and North American Industry Classification System codes. If such a project involved relating data from another state under a data sharing agreement, then the project would probably involve the need for a

translation tables that would allow data across states with different data definitions to be directly compared. The capability to translate or interpret data across systems and states has been designed into the SEAKOR model system. This example situation represents a normal research situation for a deployed SEAKOR model system.

SEAKOR Model System Integrated Training

Mason (2002) cautions that technology alone will not convert data into research, in that teacher involvement and analytic skills are necessary to convert data into research that leads to knowledge. It is important that teachers are able to manage data, ask thoughtful research questions, analyze data, and apply knowledge to classroom practice. McLeod (2005), director of the University of Minnesota School Technology Leadership Initiative, and Lachat and Smith (2005) suggest that the use of data for decision-making represents a paradigm shift for teachers using data related to classroom achievement and formative assessment of student learning. McLeod (2005) believes that teachers and instructional support staff need specialized training in order to embrace data-driven approaches. Mason (2002), reported six challenges for schools engaged in the development of data-driven decision-making processes that the SEAKOR model system and integrated training may be able to address:

1. Cultivating the desire to transform data into knowledge.
2. Focusing on a process for planned data use.
3. Committing to the acquisition and creation of data.
4. Organizing data management.
5. Developing analytical capacity.
6. Strategically applying information and results.

Group training of the type used for implementing student management systems has been found to be relatively ineffective in relation to training that provides individual interaction regarding teachers' roles in the use of data in ways that inform practice (Wayman, 2005). Regardless of role, McLeod (2005) asserts that all classroom teachers, administrators, and practitioners should be able to articulate for themselves five fundamental elements that constitute effective data-driven education:

1. Good baseline data.
2. Measurable instructional goals.
3. Frequent formative assessment.
4. Professional learning communities.
5. Focused instructional interventions.

The SEAKOR model addresses the factors reported by Mason (2002) and McLeod (2005) regarding the need to prepare teaching staff for an active role in the use of data to improve learning. The SEAKOR model integrates user-role appropriate training into all phases of the process toward preparing and using data to support pre-planned and ad hoc data-driven decisions. Integrated training for the modular SEAKOR model system accommodates the perspective of users in their roles within the process of defining research goals, the methods to accomplish these goals, the data needed for analysis, and analytical processes and reporting.

The SEAKOR model implementation training would be conducted in an experiential learning environment with a SEAKOR facilitator using role specific authentic projects and activities rather than focusing training on the use of technology, a regrettable situation that is found too often (Ronka, Lachat, Slaughter & Meltzer, 2009).

Because the SEAKOR model system consists of one software application, users with roles that change over time could find more comfort in adjusting to new roles with one software application than would be the case with more complex data warehouse systems.

Summary

Chapter VI has proposed a model for the design, development, and implementation of a longitudinal data research system that could be scaled to link similar data systems across agencies and states using relatively simple, obsolescence resistant hardware and software that may be less expensive and more capable than the types of business data warehouse systems promoted by vendors and consultants in the United States. The model system could be a fully functional system in support of best practices research. The model system could have value if used in colleges of education for appropriate teacher and administrator training.

CHAPTER VII

DISCUSSION

The U.S. government has provided almost $5 billion to states to support state longitudinal data systems (SLDS) for education and employment outcomes research. However, some recipient states, along with recognized leaders of the information technology (IT) infrastructure, and the U.S. government have reported barriers to SLDS success. If not resolved, these barriers could lead to systemic data system failure in education as has been the case with the 83.3% failure rate reported for business data systems that are typically used as the basis for educational data systems. If the more complex data systems for education mirror the failure rates of the less complex data systems used in business, this situation could constitute a disaster for American education, especially in view of the White House caution that states will be alone with funding repairs, replacements, or workable alternatives to failed educational data systems when one-time U.S. government funding support expires.

The literature has revealed some consensus on barriers to achieving success in meeting the U.S. government required elements for federally funded SLDS systems. Some states are making progress toward implementing the twelve mandatory elements (America COMPETES Act of 2007):

1. Unique statewide student identifier;
2. Student-level enrollment, demographic, and program participation information;
3. Student-level information about the points at which students exit, transfer in, transfer out, drop out, or complete P-16 education programs;
4. The capacity to communicate with higher education data systems;
5. State data audit system assessing data quality, validity, and reliability;
6. Yearly test records of individual students;
7. Information on students not tested by grade and subject;
8. A teacher identifier system with the ability to match a teacher to students;
9. Student-level transcript information, including courses completed and grades earned;
10. Student-level college readiness test scores;
11. Information on the extent to which students transition successfully from secondary school to post-secondary education, including whether students enroll in remedial course work;
12. Other information determined necessary to address alignment and adequate preparation for success in postsecondary education.

Should one assume that an educational data system that complies with the twelve federal elements be fully successful as a SLDS? Arguably, the mandatory twelve elements in themselves introduce additional issues of complexity, increasing the complexity of existing barriers to the success of SLDS. For example, full compliance with the listed twelve mandatory U.S. government requirements at any level from district level or state

level may not necessarily guarantee SLDS success at the next higher level because of the lack of standards for alignment of data definitions that allow databases to link and share related data. Alignment on this scale at the state level and interstate levels would be very expensive and does not appear to have been addressed in mandatory state or federal standards or in a separate road map that could lead to success.

The overall complexity of proposed SLDS systems would seem to reflect a recurring phenomenon in database thinking since the 1970s regarding a debate between proponents of complex models versus proponents of simpler models. More complex data systems and models seem to have won these debates more than not, and when advantages of simpler models became recognized, layers of complexity were added even though there was little performance advantage to compensate for the additional complexity. According to Hellerstein & Stonebraker (2005) history is repeating, in that many of the "...architectural innovations implemented in high-end database systems are regularly reinvented both in academia and in other areas of the software industry" (p. 1). In other words, systems recommended for use in complex education systems tend to be software and hardware re-purposed from less complex and more linear business applications that may have less hardware and software layers than education systems. For example, the U.S. government longitudinal study of student data systems reported that over 60% of school district educational data systems suffered from a lack of interoperability across multiple data systems that resulted in the inability to link data systems (America COMPETES Act of 2007). The issues that are interoperability obstacles across school districts within a state could become compounded with interoperability requirements

across state agencies and could be re-compounded with attempts to connect SLDS across state boundaries.

Adding to the barriers to SLDS and the complexity of database systems are mixed signals from the federal government that suggest the presence of a political pendulum regarding SLDS. On the one hand, the federal government has released reports criticizing the effectiveness of longitudinal data systems while at the same time allocating approximately $4.8 billion to implement them. On the one hand the U.S. Department of Education found that electronic data systems have yet to influence classroom level decision-making, on the other hand the competitive area of least emphasis in Race to the Top grant applications was the use of state data systems for improving instruction. On the one hand, there is pressure toward increased collection and use of data, while on the other hand, there are questions about the fundamental need for such complex SLDS. One the one hand, the federal government is promoting longitudinal data systems to evaluate student outcomes from higher education into the workforce, while on the other hand, the bulk of the funding is going to K-12 education. On the one hand, the federal government has enacted several laws to enforce regarding the protection of individual privacy, while on the other hand, the federal government is actively funding SLDS development grants that essentially require states to violate state and federal privacy laws in order to qualify for SLDS funding. On the one hand, while the federal government has funded approximately $4.8 billion for SLDS, recognized leaders of the IT infrastructure have publicly stated that the systems being purchased with federal funds are incapable of meeting requirements due to performance issues and obsolescence. On the one hand, the rhetoric used by the federal government to promote data warehouses speaks of education

reform, while on the other hand, approximately 80% of the funding expended on SLDS has been used to fund only hardware, software, and IT staff to manage data storage and retrieval instead of research. Could it be "We are rushing, again, into school reform initiatives with billions of dollars without much evidence that where we are headed is the right direction, and, in some cases, with evidence that it is clearly the wrong one" (Strauss. 2010; ¶ 8).

The research stress map depicted in Figure 3 describes the complexity of following individuals through multiple education systems, including multiple entry and exit points, to multiple access points related to employment and social programs. Awareness of the complexity of following individuals from P-K into secondary, postsecondary, and the workforce as illustrated in Figure 3 is becoming more apparent as existing data warehouse systems are increasingly unable to accommodate complex research needs that would seem to become more complex with each new legislative requirement related to SLDS.

As leadership models increasingly rely on the use of data to inform decisions, it is important that leaders identify and provide the means to prepare the organization to become active users of data. Successful data initiatives are replicable and scalable when they support all levels of an organization. Wayman, Midgley, and Stringfield (2006) suggest the use of collaborative teams of data users, including teams of P-20 teachers who study data for classroom improvement application. The proposed SEAKOR model research system depicted in Figure 2 was designed to respond to the literature in terms of barriers to success inherent in business data warehouse systems being re-purposed for use in education and employment outcomes research. While teachers may be critical of

accountability initiatives, they may embrace the use of data when policies are thoughtfully implemented, respond to the learning needs of students, and are considered useful in improving teaching practices. Educational data management systems designed and implemented by IT staff often ignore the perspective of teachers and other staff who have expressed frustration that data is inaccessible, and when available, the data are unusable for teacher analysis or are formatted in ways that requires further education in research and analysis.

The SEAKOR model research system is the result of a bidirectional top down—bottom up design process that starts with defining strategic and operational decision planning goals and the data that must be collected and analyzed to support them. For example, the SEAKOR model decision goal planning process shown in Figure 4 evolved from several years of applying backward planning methodology and is considered to be an essential component of designing and implementing an effective research system, especially a system designed to be operated by practitioners. This formal process, if implemented at all levels of an organization, may increase the probability of data system success while decreasing time and resources required for effective implementation. The benefit would be the result of connecting research questions to the sources of data at all levels of the education enterprise in order to support improvement in teaching and learning and employment outcomes in ways that protect the privacy of individuals.

The software application integrated into the SEAKOR model has continued to evolve in performance and adaptability since its first appearance as a commercial relational database application in 1995. The software is updated regularly with advanced features, expanded connectivity, and expanded compatibility. The system, in relation to

SLDS systems, the cost of which may exceed $10 million, could be a low cost/high benefit alternative due to less complex and less expensive hardware, software, staffing, and maintenance requirements. The SEAKOR model system is a fully functional longitudinal research database tool for conducting comprehensive research on interoperability, as well as the basic U.S. government premise that SLDS, as currently designed, could lead to measurable improvement in education. The SEAKOR model is fully capable of testing the notion that classroom teachers, administrators, and practitioners can be responsible for the quality of educational research without the intermediate control of data currently exercised by IT staff beyond data storage and retrieval. In other words, the SEAKOR system can be used to extract disaggregate raw research data from IT storage and retrieval processes from intake to selection, preprocessing, transformation, data mining, interpretation, evaluation, knowledge discovery, and reporting.

Pending large-scale acceptance of research data systems such as the SEAKOR model, perhaps classroom teachers and practitioners could earn professional development credit through a college of education as the result of demonstrating competence in the use of databases for classroom integration, educational research, and data-driven decision-making. Experiential learning laboratory workshops in database technology could initiate such a process. Were colleges of education to embrace the notion of database technology preparation in pre-service or master's degree programs, it may be possible to increase the level of informed constructive criticism of IT centered practices in education to the point of creating a catalyst for a paradigm shift in education that would help the education community to regain control of the processes for which they are responsible. Introducing

the use of database technology as an educational tool could facilitate such a process at a time when teachers and other education practitioners are feeling threatened by the proponents of SLDS and standardized testing movements who seem to justify them with the promise that SLDS and standardized testing will be used to weed out bad teachers using database statistical methods known to have a failure rate of 83.3%.

This dissertation has involved extreme effort in searching for a more balanced appraisal of the state of SLDS effectiveness. However, positive information beyond white papers provided by technology vendors could not be located. That said, an area for further research could be related to examining U.S. and state government formal proof of concept for SLDS design criteria regarding implementation interconnectivity within and across states. In other words, this dissertation could not identify the existence of authoritative research that would offer confidence that the SLDS movement will result in much beyond great expense of one time federal funding that places a burden on states for perpetual maintenance and support.

Another area for further research could be to evaluate the consequences and possible remedies regarding difficulties that individuals face when incorrect personal and education data are released, legally or illegally, to the National Student Clearinghouse (NSC) and other databases for public, educational, and employment scrutiny. Anecdotal evidence exists regarding the lack of respect for student record privacy, in that organizations that legally or illegally receive student privacy data may distribute such data freely without consequences.

There also exist anecdotal reports of inaccurate information in the NSC. For example, perspective employers may query the NSC for evidence that supports a

professional position applicant's claim that a degree was earned at a specific institution in a certain year. If data provided to the NSC was incorrect or missing, the job applicant's opportunity could be lost without the issues being made known to the individual whose data was mishandled. During the research related to this dissertation, no remedy was discovered that would allow individuals to block or view data provided by educational institutions to entities such as the NSC or to correct inaccuracies in data resulting from incorrect data entry or subsequent processing. The potential problem would be compounded as student data containing privacy information begins to cross state boundaries.

Finally, the SEAKOR model could be prototyped using a building block, modular approach that would start with the development of a localized research stress map, comprehensive data dictionary, and decision goal planners to support SEAKOR implementation initially for one PK-12 school, one postsecondary institution, and their related agencies to build a success oriented implementation plan designed to eliminate the issues that arise when complex systems are implemented without proper preparation. As an interim system, the SEAKOR model could be used to identify and resolve issues involving connectivity, sustainability, scalability, security, usability, and training at every level in preparation for the implementation of SLDS on the scale envisioned by the U.S. government and states who have accepted one-time SLDS grants.

Appendix

GLOSSARY

Accountability - the state of being accountable, liable, or answerable; In education, a policy of holding schools and teachers accountable for students' academic progress by linking such progress with funding for salaries, maintenance, etc.

Aggregation - the combining of two or more objects into a single object; combining the results of all groups that make up the sample or population.

Analysis - the examination of facts and data to provide a basis for effective decisions.

Common student identifier - a grouping of numbers and/or letters associated with only one student used to identify and follow student data through K-12, postsecondary, and into the workforce.

Commercial off-the-shelf (COTS) - a term defining computer software or hardware systems, that could present limitations but that are ready-made and available for free, lease, sale, or license to the general public.

Connectivity - ability to make and maintain a connection between two or more points within a computing system.

Container field - a database field configured to store files that may be graphics, images, documents, spreadsheets, and portfolios. A container field may be accessed by double clicking on the file icon shown in the container field.

Data analysis - the process of evaluating data using analytical and logical reasoning to reach a finding or conclusion.

Database - a collection of data for one or more uses. Databases store data in accessible fields. Data fields include text, numbers, dates, summaries, calculations, and containers.

Database architecture - describes how data are processed, stored, and utilized in a data system.

Database field - the basic unit of data entry. In comparison to a spreadsheet where a record is the data contained in a row, a field is the data contained in a column, such as name or birth date.

Database tuning - various procedures used to optimize the performance of a database.

Data definition - factual information used as the basis for reasoning and calculation.

Data dictionary - a collection of descriptions of data as objects contained in data fields. A data dictionary entry for name could be the criteria of last name or full name.

Data-driven decision making (DDDM) - making decisions based on demographic, student learning, perceptions, and school process data. True data-driven decision-making has at the center of every decision the guiding principles of the learning organization.

Data extract - a copy of data output from a database as of a specific day and time generally for use in a database for analysis. Enrollment data extracted on September 15 at noon would contain only the data up to that time. In other words, a data extract will not automatically update.

Data integration - involves combining data from multiple sources in a record for processing. An example could be data related to programs of study with the definition of the classification of instructional program included in the record from another file.

Data Mart - a database or data table containing data limited to a specific purpose or subject. An example could be enrollment records.

Data mining - a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data; the process of automatically discovering useful information in large data repositories in order to find novel and useful patterns that might otherwise remain unknown; techniques for finding patterns and trends in large data sets; the process of automatically extracting valid, useful, previously unknown, and ultimately comprehensible information from large databases.

Data quality - information that meets the requirements of its authors, users, and administrators

Data translation - a pre-processing step in assembling data for analysis that established common definitions for similar data coming from different sources. An example could be a file with hyphenated social security numbers being combined with a file with social security numbers in a different format. A translation calculation reforms records to meet a standard definition.

Data warehouse - a collection of data from various sources that is pre-processed and validated before being used for analytical procedure.

Design-build - in contrast to commercial off-the-shelf (COTS) hardware and software procurement practices typically used in information technology system development that may result in multiple software and hardware systems with compatibility issues, the design-build philosophy combines architecture/engineering and construction in one

contract under one entity. This type of project partnership may reduce time, save money, provide stronger guarantees, and produce obsolescence-resistant results.

Disaggregate data - data that are not summarized or sub-summarized into groupings that omit the personal identifier attached to each record. Disaggregate data are defined as personally identifiable data.

Domain knowledge - internalized knowledge that is difficult to transfer to another individual verbally or in writing. Domain knowledge is situational recall.

Encryption - the process of transforming information or data through mathematical algorithms in order to render information or data unreadable to anyone who does not have knowledge to decrypt or make the information readable. The process generally involves using a reverse algorithm called a key.

Gigabyte - a multiple of the unit of digital information storage called a byte. A gigabyte is 1 billion bytes, abbreviated as GB. A gigabyte is 1×10^9.

Hardware - anything that represents the physical components of a computer system, such as hard drive, keyboard, and monitor.

Hardware key - a physical encryption device that contains the ability to encrypt data stored within a computing system that cannot be accessed or read without insertion of a key.

Information Technology (IT) - a set of tools, processes, and methodologies such as programming data conversion, storage, and retrieval necessary to collect, process, and present information.

Interface - equipment or programs designed to communicate information from one system of computing devises or programs to another.

Interoperability - the ability of computer systems to work together. Computer makers create design features that allow the output of one computer to be understood by a different computer and vice versa.

IT - Information technology.

IT professionals - an occupational field engaged in the study, design, development, implementation, support, or management of information systems, a field that is separate from computer science and information systems.

Knowledge discovery in databases (KDD) - overall process of converting raw data into useful information.

Knowledge management (KM) - comprises a range of strategies and practices used in an organization to identify, create, represent, distribute, and enable adoption of insights and experiences. Such insights and experiences comprise knowledge, either embodied in individuals or embedded in organizational processes or practice.

Mainframe computer - a term used to describe large scale computer systems designed in the 1950s and 1960s when computer hardware components were physically large in size. A computer with the capabilities of a typical desktop computer in the early years could occupy 1,000 square feet.

Megabyte - a multiple of the unit of digital information storage called a byte. A megabyte is 1 million bytes, abbreviated as Mb. A megabyte is 1×10^6.

Metadata - generally considered to represent data on the structure of data, that is data that is used to describe data definitions, structures, and administration.

Persistent data - data that remains unchanged until revised.

Pre-processing - procedures performed on raw data in preparation for analysis. Pre-processing includes data validation and transformation.

Personally identifiably student data - information that can be used to locate or identify an individual through name, alias, social security number, biometric record, or other information that may lead to identify theft of fraudulent use of information that may result in substantial harm, embarrassment, and/or inconvenience to individuals.

Query - specifications of procedures communicated to a computer that provide instructions for processing data, such as finding records that were created on a certain date or finding records labeled with the same last name along with how the found records should be displayed.

Related data/relationship - in order for the data of records in multiple databases to be combined for analysis, a database must be instructed how to match records across tables or databases. The process is called creating a relationship. An example of a relationship might be a classification of instructional program code in one database that is matched with a federal database in order to import the description of an instructional program that matches the code.

Relational database management system - two or more database tables joined through a matching identifier in a common field that allows the information contained in the multiple tables to be accessed and analyzed as one set of data. Student identification numbers, social security numbers, and name-birth date strings are examples of identifiers that can be used to link data tables.

Scalability - how well a solution to some problem will work when the size of the problem is increased or decreased.

SEA [state education agency] - is a formal governmental label for the state-level government agencies within each U.S. state responsible for providing information, resources, and technical assistance on educational matters to schools and residents.

Server - a computer that links a series of computers into a system. An example would be connecting all computers on a college campus to a special server computer that directs the flow of communications throughout the system.

Software - a collection of computer programs that provide instructions for computer processes.

SQL (Structured Query Language) -

State Longitudinal Data System (SLDS) - intended to enhance the ability of States to efficiently and accurately manage, analyze, and use education data, including individual student records.

Sub-aggregate - a level of combining data records. A database may contain multiple records of enrollments for one student over a period of four years. A sub-aggregated record could combine the enrollments by semester, by year, or classification.

Terabyte - a multiple of the unit of digital information storage called a byte. A terabyte is 1 trillion bytes, abbreviated as TB. A terabyte is 1 x 1012.

Visualization - consists of methods that communicate information through graphical means. An example could be a logic tree, graph, circuit wiring diagram, or a block diagram description of a data warehouse.

REFERENCES

Agrawal, V. (2005). Data warehouse operational design: View selection and performance simulation (Doctoral dissertation, University of Toledo, 2005). *Dissertation Abstracts International*, 66, 1417.

Aikin, W. M. (1942). Foreword. In E. R. Smith & R. W. Tyler, *Adventure in American education volume III: Appraising and recording student progress*. New York: McGraw-Hill.

Alves, C., & Finkelstein, A. (2002). Challenges in COTS decision-making: A goal-driven requirements engineering perspective. *Proceedings of the 14th international conference on software engineering and knowledge engineering, Italy, 27, 789-794.*

America COMPETES Act of 2007, Pub. L. No. 110-69 § 6401 (a)(1) (2007).

America Recovery and Reinvestment Act of 2009, Pub. L. No. 111-5 § 14005 (2009).

Bakst, D. (2009). State data systems: Privacy and security issues should trump the need for data. *Council on Law in Higher Education*. Retrieved May 20, 2010, from http://www.clhe.org/marketplaceofideas/data-security/state-data-systems-privacy-and-security-issues-should-trump-the-need-for-data/

Banta, T. W. (2007). A warning on measuring learning outcomes. *Inside Higher Education*. Retrieved August 15, 2010, from http://www.insidehighered.com/views/2007/01/26/banta

Basken, P. (2010, January 3). States embrace student-data tracking, with prodding from White House. *The Chronicle of Higher Education*. Retrieved August 2, 2010, from http://chronicle.com/article/States-Embrace-Student-Data/63376/?sid=at

Betts, M. (2003). Why IT projects fail: There's a difference between managing risk and preventing failure. *Computerworld*. Retrieved August 29, 2010, from http://www.computerworld.com/s/article/84266/Why_IT_projects_fail

Boccaletti, S., Grebogi, C., Lai, Y.-C., Mancini, H., & Maza, D. (2000). The control of chaos: Theory and application. *Physics Reports*, 329(2000), 103-197.

Bodamer, R. (2010). Time for program managers to embrace agile development. *Washington technology*. Retrieved July 21, 2010, from http://washington technology.com/articles/2010/01/22/program-managers-need-to-become-agile.aspx

Brachman, R. J., & Anand, T. (1996). The process of knowledge discovery in databases. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 10-20). Menlo Park, CA: AAAI Press.

Brooks, F. P. (1987). No silver bullet: Essence and accidents of software engineering. *Computer*, 20(4), 10-19.

Burke, J. C. (2004). *Accountability reporting with so much effort; why so little effect*? Testimony to the National Commission on Accountability in Higher Education, 2004. Retrieved August 14, 2010, from http://www.sheeo.org/account/comm/testim/Burke testimony.pdf

Carey, K., & Aldeman, C. (2008). *Education sector reports: Ready to assemble: A model state higher education accountability system*. Washington, D.C.: Education Sector.

Center on Law and Information Policy. (2009). *Children's educational records and privacy: A study of elementary and secondary school state reporting systems*. Retrieved October 29, 2009, from http://law.fordham.edu/assets/CLIP/CLIP_Report_Childrens_Privacy_Final.pdf

Charette, R. (2006). Why software fails. *IEEE Spectrum*. Retrieved July 24, 2010, from http://www.spectrum.ieee.org/computing/software/why-software-fails

Charette, R. (2010). US government identifies 26 high-risk IT projects for intense reviews. *IEEE Spectrum: Inside Technology*. Retrieved September 5, 2010, from http://spectrum.ieee.org/riskfactor/computing/it/us-government-identifies-26-highrisk-it-projects-for-intense-reviews

Children Now. (2009). *California policy brief: Children's Education: The clear case for data systems redesign, California*. Author. Retrieved August 6, 2010, from http://www.hewlett.org/uploads/files/ChildrenNow_ChildrensEducation.pdf

Cios, K. J., Pedrycz, W., Swiniarski, R. W., & Kurgan, L. A. (2007). *Data Mining: A knowledge discovery approach*. New York: Springer.

Confidential Information Protection and Statistical Efficiency Act, 31 U.S.C. 1104(d); 44 U.S.C. 3504 (2002).

Data Management Association. (2009). *DAMA guide to the data management body of knowledge* (1st ed.). Bradley Beach, NJ: Author.

Data Quality Campaign. (2009). *ARRA support for state longitudinal data systems*. Retrieved August 28, 2010, from http://www.dataqualitycampaign.org/files/ARRA_Support_for_SLDS_-_FINAL.pdf

Deming, W.E. (2000). *Out of the crisis*. Boston: MIT Press.

Department of Education. (2009). *The American Recovery and Reinvestment Act of 2009: Education jobs and reform*. Retrieved August 28, 2010, from http://www2.ed.gov/print/policy/gen/leg/recovery/factsheet/overview.html

Design-build Institute of America. (2010). *What is design-build?* Retrieved July 25, 2010, from http://www.dbia.org/about/designbuild/default.htm?PF=1

Dewitt, D. J., & Gray, J. (1992). Parallel database systems: The future of high performance database processing. *Communications of the ACM*, 36(6) 1-26.

Education Commission of the States. (1998). *Learning and technology: Integrating policy perspectives and research*. Denver, CO: Author.

Education Technical Assistance Act of 2002 Public Law 107-79 (2002).

Evans, M. (2005). Overdue and over budget, over and over again. *Economist*. Retrieved July 10, 2010 from http://flyvbjerg.plan.aau.dk/News in English/Economist 090605Traffic.pdf

Family Educational Rights and Privacy Act, 20 U.S.C. § 1232g; 34 CFR Part 99 (1976).

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From data mining to knowledge discovery in databases. *American Association for Artificial Intelligence*. Retrieved October 9, 2009, from https://www.aaai.org/aitopics/assets/PDF/AIMag17-03-2-article.pdf

Federal Bureau of Investigation. (2005). *The Federal Bureau of Investigation's management of the trilogy information technology modernization project* (Office of the Inspector General Audit Report No. 05-07). Washington, D.C. Retrieved September 17, 2010, from http://www.justice.gov/oig/reports/FBI/a0507/exec.htm

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). *Knowledge discovery in databases: An overview*. Retrieved October 10, 2009, from https://www.aaai.org/ojs/index.php/aimagazine/article/viewPDFInterstitial/1011/929

Goldstein, H. (2005). Who killed the virtual case file: How the FBI blew more than $100 million on case-management software it will never use. *IEEE Spectrum: Inside Technology*. Retrieved September 5, 2010, from http://spectrum.ieee.org/computing/software/who-killed-the-virtual-case-file/0

Health Care and Education Reconciliation Act, PL 111-152, (2010).

Health Insurance Portability and Accountability Act, PL 104-191, (1996).

Hellerstein, J. M., & Stonebraker, M. (2005). Anatomy of a database system. In J. M. Hellerstein & M. Stonebraker (Eds.), *Readings in database systems* (4th ed.). (pp. 1-54). Cambridge, MA: MIT Press.

Henry, P. (2007). The case against standardized testing. *Minnesota English Journal*, 43(1), 39-71. Retrieved August 13, 2010, from http://www.mcte.org/journal/mej07/mej07.pdf

Higher Education Opportunity Act of 2008, Pub. L. No. 110-315 § 113 (2008).

Idaho Legislature: Office of Performance Evaluations. (2006). *Idaho student information management system (ISIMS): Lessons for future technology projects*. Retrieved January 10, 2010, from http://www.legislature.idaho.gov/ope/publications/reports/r0602.htm

Kanaracus, C. (2010). *Forrester: 2010 IT spending still looks strong*. Forrester Research CIO. Retrieved July 25, 2010, from http://www.cio.com/article/600114/forrester_2010_IT_Spending_Still_Looks_Strong

Khabaza, T. (2009). *Hard hat area: Myths and pitfalls of data mining*. Chicago, IL: SPSS. Retrieved July 27, 2010, from http://searchcio-midmarket.bitpipe.com/detail/RES/1235145316_732.html

Khandani, S. (2005). *Industry initiatives for science and math education (IISME): Engineering design process*. Retrieved July 25, 2010, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.118.5172&rep=rep1&type=pdf

Kline, J. (2010). Letter from John Kline, Senior Republican member: Committee on Education and Labor: U.S. House of Representatives to The Honorable Arne Duncan, Secretary, U.S. Department of Education. Dated February 23, 2010. Retrieved August 28, 2010, from http://law.fordham.edu/center-on-law-and-information-policy/14769.htm

Kowalski, T. J., Lasley, T. J., & Mahoney, J. W. (2008). *Data-driven decisions in school leadership: Best practices for school improvement*. Boston: Allyn & Bacon.

Krigman, E. (2009). Are state data systems worth the risk? *National Journal Online*. Retrieved July 21, 2010, from http://education.nationaljournal.com/2009/10/are-student-data-systems-worth.php?rss=1

Lachat, M. A., & Smith, S. (2005). Practices that support data use in urban high schools. *Journal of Education for Students Placed at Risk*, 10(3), 333-349.

Laird, E. (2008). *Developing and supporting P-20 education data systems: Different states, different models*. Retrieved April 30, 2010, from http://www.dataquality campaign.org/ resources/92

Lai, E. (2007). Relational database pioneer says technology is obsolete. *Computerworld.* Retrieved from July 22, 2010, from http://www.computerworld.com/s/ article/print/ 9034619/Relational_database_pioneer_says_technology_is_obsolete

Law, Y-N., Wang, H., & Zaniolo, C. (2004). Proceedings of the thirtieth international conference on very large data bases: Vol. 30. *Query languages and data models for database sequences and data streams* (pp. 492-503). Toronto: VLDB Endowment.

Lederman, D. (2009, September 28). Aggressive plan for state data systems. *Inside Higher Education.* Retrieved July 22, 2010, from http://www.insidehighered.com/ news/2009/09/28/data

Lederman, D. (2010, February 1). Clash over student privacy. *Inside Higher Education.* Retrieved August 2, 2010, from http://www.insidehighered.com/layout/ set/print/news/2010/02/01/ferpa

Lederman, D. (2010, May 12). Using data to drive performance. *Inside Higher Education.* Retrieved August 3, 2010, from http://www.insidehighered.com/ layout/set/print/news/focus/assessment_and_accountability/recent/analytics

Luan, J. (2002a). Data mining and its applications in higher education. In A. M. Serban, & J. Luan. (Eds), *New directions for institutional research: Knowledge management: Building a competitive advantage in higher education* (pp. 17-36). San Francisco.

Luan, J. (2002b). *Data mining and knowledge management in higher education: Potential application [Abstract]*. Paper presented at the annual forum for the Association for Institutional Researchers, Toronto, Ontario, Canada.

Luan, J. (2006). *Data mining applications in higher education.* Chicago, IL: SPSS. Retrieved July 27, 2010, from http://www.spss.ch/upload/1122641492_Data mining applications in higher education.pdf

Luan, J., & Willett, M. S. (2001). *Data mining and knowledge management*. Paper presented at the annual conference of the Association for Institutional Researchers, Long Beach, CA.

Mandinach, E. B., Honey, M., & Light, D. (2006). *A theoretical Framework for data-driven decision making*. Paper presented at the annual meeting of American Education Research Association, San Francisco. Retrieved July 20, 2010, from http://cct.edc.org/admin/publications/speeches/DataFrame_AERA06.pdf

Marzano, R. (2003). Using data: Two wrongs and a right. *Educational Leadership*, 60(5), 56-60.

Mason, S. (2002). *Turning data into knowledge: Lessons from six Milwaukee public schools*. Paper presented at the annual conference of the American education research association, New Orleans, LA.

Mazon, J. N., Pardillo, J., & Trujillo, J. (2007). A model-driven goal-oriented requirement engineering approach for data warehouses. In J. L. Hainaut, et al. (Eds.), *Proceedings of the 2007 Conference on Advances in Conceptual Modeling: Foundations and Applications* (pp. 155-264). Berlin: Springer-Verlag.

McDonald, S., Andal, J., Brown, K., & Schneider, B. (2007). *Getting the evidence for evidence-based initiatives: How the Midwest states use data systems to improve education processes and outcomes* (Issues & Answers Report, REL 2007—No. 016). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Midwest. Retrieved March 22, 2010, from http://ies.ed.gov/ncee/edlabs/projects/project.asp?ProjectID=29

McKenna, R. (2009). Memorandum from Rob McKenna, Washington State Attorney General to Fellow State Attorneys General. Dated December 3, 2009. Retrieved August 28, 2010, from http://law.fordham.edu/center-on-law-and-information-policy/14769.htm

McLeod, S. (2005). Data-driven teachers. *School Technology Leadership Initiative*. Retrieved July 10, 2010, from http://download.microsoft.com/download/2/5/9/259f7395-bd6a-45d0-bbe2-cb7cbc3e16a7/ThoughtLeaders_DDDM_May05.doc

Merola, L. (2006). The COTS software obsolescence threat. *Proceedings of the Fifth International Conference on Commercial-off-the Shelf (COTS)-Based Software System.* USA, 1-7.

Mills, J. (2005). State data systems and privacy concerns: Strategies for balancing public interests. *Jobs for the Future*. Retrieved July 26, 2010, from http://www.jff.org/publications/education/state-data-systems-and-privacy-concerns-/326

Molenaar, K. R., Songer, A. D., & Barash, M. (1999). Public-sector design/build evolution and performance. *Journal of Management in Engineering*, 15(2), 54-62.

Mohan, R. (2006). [Letter to members, Joint Legislative Oversight Committee, Idaho Legislature]. In Office of performance evaluations. *Idaho student information management system (ISIMS): Lessons for future technology projects* (pp. iii). Boise, Idaho: State of Idaho.

Monash, C. A. (2006). Bulletins from the database management front. *Computerworld.* Retrieved July 22, 2010, from http://www.computerworld.com/s/article/111319/Bulletins_From_the_Database_Management_Front

Mosley, M. (2008). *DAMA-DMBOK functional framework, Version 3.02.* Lutz, FL: Data Management Association International. Retrieved July 27, 2010, from http://www.dama.org/i4a/forms/form.cfm?id=29

Mullin, C., & Lebesch, A. (2010). *Moving success from the shadows: Data systems that link education and workforce outcomes.* Retrieved June 8, 2010 from http://www.aacc.nche.edu/Publications/Briefs/Documents/successshadows_03162010.pdf

National Commission on Accountability in Higher Education. (2005). *Accountability for better results: A national imperative for higher education.* Boulder, Colorado: Author. Retrieved August 6, 2010, from http://www.sheeo.org/account/accountability.pdf

National Student Clearinghouse. (2010). *National Student Clearing house: The nation's trusted source for student degree and enrollment verification.* Retrieved July 26, 2010, from http://www.studentclearinghouse.org

Nemati, H. R., Steiger, D. M., Iyer, L. S., & Herschel, R. T. (2002). Knowledge warehouse: An architectural integration of knowledge management, decision support, artificial intelligence and data warehousing. *Decision Support Systems,* 33(2), 143-161. Retrieved September 10, 2010, from http://linkinghub.elsevier.com/retrieve/pii/s0167923601001415

Northwest Environmental Data Network. (2006). *Best practices for data dictionary definitions and usage.* Portland, Oregon: Author. Retrieved August 6, 2010, from http://www.nwcouncil.org/ned/DataDictionary.pdf

Okes, D. (2009). *Root cause analysis: The core of problem solving and corrective action.* Milwaukee: Quality Press.

Palaich, R. M., Good, D. G., & van der Ploeg, A. (2004). *Research-based analysis of education policy: Policy issues: State education data systems that increase learning and improve accountability No.16.* ERIC ED 489522. Retrieved October 15, 2009, from http://eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED489522

Paley, A. R. (2007, April 16). Data-mining of students raises alarms officials might halt access to load database. *Washington Post.* Retrieved May 2, 2010, from http://www.boston.com/news/education/higher/articles/2007/04/16/data_mining_of_students_raises_alarms/

Port, D., & Chen, S. (2004). Assessing COTS assessment: How much is enough? In R. Kazman & D. Port (Eds.), *Lecture notes in computer science* (pp. 183-198). Springer: Heidelberg, Germany: Springer Verlag.

Prescott, B. T., & Ewell, P. (2009). *A framework for a multi-state human capital development data system.* Retrieved March 29, 2010, from http://www.wiche.edu/info/publications/FrameworkForAMultistateHumanCapitalDevelopmentDataSystem.pdf

Privacy Rights Clearinghouse. (2010). *Filtered data breaches*. Retrieved July 23, 2010, from http://www.privacyrights.org/data-breach

Protection of Pupil Rights Amendment 20 U.S.C. § 1232h; 34 CFR Part 98 (1978).

Romero, C., & Ventura, S. (2007). Educational data mining: A survey form 1995-2005. *Expert Systems with Applications* 33(2007), 135-146.

Ronka, D., Lachat, M. A., Slaughter, R., & Meltzer, J. (2009, January). Answering the questions that count. *Educational Leadership*, 66(4), 18-24.

Sandborn, P. (2007). Software obsolescence: Complicating the part and technology obsolescence management problem. *IEEE Transactions on Components and Packaging Technologies,* 30(4), 886-888.

Schneider, H. (2005). *Rapid ICT change and workplace knowledge obsolescence: Causes and proposed solutions (Research Publication No. 2005-04)*. Cambridge, Massachusetts: Harvard University, Berkman Center for Internet & Society. Retrieved July 29, 2010, from http://cyber.law.harvard.edu/publications

Sen, A., & Sinha, A. (2005). A comparison of data warehousing methodologies: Using a common set of attributes to determine which methodology to use in a particular data warehousing project. *Communications of the ACM*, 18(3), 79-84.

Serban, A. M., & Luan, J. (2002). Overview of knowledge management. In A. M. Serban & J. Luan (Eds.), *New directions for institutional research: knowledge management: Building a competitive advantage in higher education* (pp. 5-16). San Francisco: Jossey-Bass.

Shiley, C. S. (2003). *Putting the rights into the Family Education Rights and Privacy Act: Enforcement and the private right of action.* Bachelor's thesis, Massachusetts Institute of Technology, Cambridge.

Smith, P. (2004). *The quiet crisis*. San Francisco, CA: Jossey-Bass.

Stilton, P. (2010). Education Reform Part I: Data based decision making can help schools identify and correct problems. *Jackson News*. Retrieved March 21, 2010, from http://www.jacksonnjonline.com/2010/03/21/education-reform-part-i-data-based-decision-making-can-help-schools-identify-and-correct-problems/

Stonebraker, M. (2007a). *In response to Monash's post on the four categories of RDBMS.* Retrieved July 30, 2010, from http://databasecolumn.vertica.com/index.php?s=the+four+categories+of+RDBMS

Stonebraker, M. (2007b). *Just in time decompression in analytic databases*. Retrieved July 30, 2010, from http://databasecolumn.vertica.com/database-innovation/just-in-time-decompression-in-analytic-databases/

Stonebraker, M. (2007c). *Stonebraker comments on OODB market failures, data warehousing pain, and column advantages*. Retrieved July 30, 2010, from http://databasecolumn.vertica.com/database-architecture/stonebraker-comments-on-oodb-market-failures-data-warehouse-pain-and-column-advantages/

Stonebraker, M. (2007d). *The truth about MPP & Data warehousing*. Retrieved July 30, 2010, from http://databasecolumn.vertica.com/database-architecture/the-truth-about-mpp-data-warehousing/

Strauss, V. (2010, June 1). The desperation of race to the top. *Washington Post*. Retrieved August 2, 2010, from http://voices.washingtonpost.com/answer-sheet/education-secretary-duncan/the-desperation-of-race-to-the.html

Student Aid and Fiscal Responsibility Act of 2009, Pub. L. No 111-152 (2010).

Student Aid and Fiscal Responsibility Act (2010). H.R. 4872 RH, 111th Cong. (2010).

Tiwana, A. (2002). *The knowledge management toolkit: Orchestrating IT, strategy, and knowledge platforms* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

U.S. Bureau of Labor Statistics. (2010). *Confidentiality of data collected by BLS for statistical purposes, Washington, DC.* Author. Retrieved July 21, 2010, from http://data.bls.gov/bls/confidentiality.htm

U.S. Department of Education. (2009a). *Application for grants under the statewide longitudinal data system recovery act grants: CFDA #84.384A: PR/Award # R384A100012*. Washington D.C

U.S Department of Education. (2009b). *Race to the top program executive summary. Washington, DC*: Author. Retrieved August 29, 2010, from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

U.S. Department of Education. (2010). *Family Educational Rights and Privacy Act (FERPA), Washington, DC*: Author. Retrieved July 21, 2010, from http://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html

U.S. Department of Education, Office of Planning, Evaluation, & Policy Development. (2010). *Use of education data at the local level from accountability to instructional improvement, Washington, DC*: Author. Retrieved August 3, 2010, from http://www.ed.gov/about/offices/list/opepd/ppss/reports.html#edtech

U.S. Federal Trade Commission. (2007). Federal trade commission: Protecting America's consumers. Retrieved August 15, 2010, from http://www.ftc.gov/report/privacy3/priv-23.shtm

U.S. Government Accountability Office. (2008). *Information security: Protecting personally identifiable information, Washington, DC.* Author. Retrieved May 15, 2010, from http://www.dodig.mil/fo/Privacy/PDFs/GAOReportProtecting PII(GAO-08-343).pdf

Wayman, J. (2005). Involving teachers in data-driven decision making: Using computer data systems to support teacher inquiry and reflection. *Journal of Education for Students Placed at Risk*, 10(3), 295-308.

Wayman, J. C., Midgley, S., & Stringfield, S. (2006). *Leadership for data-based decision-making: Collaborative educator teams*. Paper presented at the 2006 Annual Meeting of the American Educational Research Association, San Francisco. Retrieved August 5, 2010, from http://edadmin.edb.utexas.edu/datause/

Wayman, J. C., Stringfield, S., & Yakimowski, M. (2004). *Software enabling school improvement through analysis of student data (Report No. 67).* Baltimore, Maryland: Center for Research on the Education of Students Placed at Risk. Retrieved August 3, 2010, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.135.6556&rep=rep1&type=pdf

Weinberger, J. (2004). Technology never fails, but projects can. *Baseline Magazine*. Retrieved July 22, 2010, from http://www.baselinemag.com/c/a/Projects-Management/Technology-Never-Fails-But-Projects-Can/

White House. (2009). *American Recovery and Reinvestment Act: The largest investment in education in our nation's history — to prevent teacher layoffs, making key education improvements and help make college affordable*. Retrieved August 28, 2010, from http://www.whitehouse.gov/assets/documents/Recovery_Act_Education_2-17.pdf