Chapter 4
A Case Study on Data Quality, Privacy, and Entity Resolution

William Decker
University of Arkansas – Little Rock, USA

John Talburt
University of Arkansas – Little Rock, USA

Fan Liu
University of Arkansas – Little Rock, USA

Pei Wang
University of Arkansas – Little Rock, USA

Ningning Wu
University of Arkansas – Little Rock, USA

ABSTRACT

This chapter presents ongoing research conducted through collaboration between the University of Arkansas at Little Rock and the Arkansas Department of Education to develop an entity resolution and identity management system. The process includes a multi-phase approach consisting of data-quality analysis, selection of entity-identity attributes for entity resolution, development of a truth-set, and implementation and benchmarking of an entity-resolution rule set using the open source entity-resolution system named OYSTER. The research is the first known of its kind to evaluate privacy-enhancing, entity-resolution rule sets in a state education agency.

INTRODUCTION

Pending regulation and recent funding opportunities present a unique opportunity to evaluate data housed in academic organizations and to improve processes and frameworks through which data systems may be improved. Research yields a number of instances where student academic data has been breached: one institution had records for 40,000 students compromised in 2010 (University of Hawaii, 2010); another institution reported Social Security numbers (SSNs), along with confidential student academic records, were breached for 21,000 students (Miami University, 2005).

Research for an entity resolution (ER) framework, which relies less on confidential student data, is described in this case study. This approach is characterized as context-sensitive, as it utilizes data elements considered "directory information." In this approach, we define directory information as data attributes about a student that can be

DOI: 10.4018/978-1-4666-4892-0.ch004

made legally public. In the United States (U.S.) education system, this is regulated by the Family Educational Rights and Privacy Act (FERPA) under the U.S. Department of Education. FERPA-based identity attributes can utilize information commonly published in telephone directories and might include attributes such as student name, address, telephone, date of birth, field of study, and grade-level, among other low-risk identifiers.

Motivated by these challenges, research is described in this case study to accomplish three primary research objectives: 1) data quality analysis of educational data; 2) development of a truth set and a privacy-preserving entity resolution (ER) framework; and 3) a quantitative evaluation of the ER framework's value. The work is motivated by both ongoing challenges inherit in academic databases and the requirement to correctly resolve student records while preserving the privacy of student data.

BACKGROUND

Public elementary and secondary schools had 49.4 million students enrolled in the 2009-10 school year (Institute of Education Sciences, 2011). In the same year, approximately 21.5 million students were enrolled in public post-secondary schools. Institutions receiving public funds from the U.S. Department of Education (USDOE) are subject to the FERPA (20 U.S.C. § 1232g; 34 CFR Part 99). FERPA is administered by the Family Policy Compliance Office of the U.S. Department of Education.

Under the precept of better decisions require better information, significant resources were provided to develop databases that house educational data and empower educators with data-driven decision-making capabilities. This is evidenced through the USDOE's award of four rounds of statewide longitudinal data system (SLDS) grants to U.S. states in 2006, 2007 and 2009 (two rounds were issued in 2009). Forty-one states and the District of Columbia received at least one grant to build a statewide longitudinal data system (Institute of Education Sciences, 2011). These systems were built in order to "efficiently and accurately manage, analyze, and use education data, including individual student records" (Institute of Education Sciences, 2011).

One need not extensively search to find evidence of student privacy breaches. In 2005, a former University of Texas at Austin student was found guilty of hacking the university system and "illegally possessing almost 40,000 Social Security numbers" (Associated Press, 2005). In October 2010, the University of Hawaii-West Oahu released a statement notifying approximately 40,000 individuals that their personal information may have been compromised (University of Hawaii, 2010). In September 2012, Miami University reported, "a grade report from the fall 2002 semester had been unwittingly placed by a now-retired faculty member into a file that was accessible via the Internet. The report included the Social Security numbers and grade information on the more than 21,000 students" (Miami University, 2005). In a 2006 issue of EDUCAUSE Review, it was reported that colleges and universities accounted for more than one-third of the publicly reported information security breaches in 2005 and the first half of 2006, yet colleges and universities provide insufficient training in privacy and security issues (Cate, 2006).

In an effort to help address these challenges, the Arkansas Department of Education (ADE) started collaborations with the Center for Advanced Research in Entity Resolution and Information Quality (ERIQ) at the University of Arkansas at Little Rock (UALR) in 2010. The OYSTER Open Source Project, which was first used as a tool for teaching entity resolution at UALR, was initially supported by much of this work with the ADE. OYSTER has since developed into a platform capable of supporting record linking, entity