

TrustEd: A Dual-Database Trusted Broker System for Sharing Multi-Agency Data

Neal Gibson
Arkansas Research Center

Abstract:

The state of Arkansas has created a P20W longitudinal data system that links data between the Arkansas Department of Education, Arkansas Department of Higher Education, Arkansas Department of Human Services, Arkansas Department of Career Education, Arkansas Department of Workforce Services, and the University of Arkansas for Medical Sciences. To protect individual privacy and comply with state and federal requirements, these data are stored in a dual-database system which keeps personally identifiable data separate from the data of research interest. A trusted broker is used to link disparate agency data together as needed for research. As many states are in the process of creating their own such “cradle to grave” systems, it is our belief that states should follow a similar path for the complete protection of individual privacy in the creation of such systems.

TrustEd: A Dual-Database Trusted Broker System for Sharing Multi-Agency Data

Neal Gibson
Arkansas Research Center

As states build data systems that link data between multiple agencies, many privacy and regulatory concerns are raised. In addition to the possible leaks of personally identifiable information (PII), another major concern is the uses of such “cradle to grave” systems, which could be mined for a wealth of information beyond the policy questions they are designed to answer. To alleviate these concerns, the state of Arkansas has created a dual-database system which keeps PII separate from data used for research. Research data is stored in a de-identified state, with records for each agency using a unique, agency specific ID that has no direct link to any other agency’s data. When research data is needed that requires the linking of two or more agencies’ data, a temporary crosswalk is used for the link, and this crosswalk is destroyed immediately after the result set is created.

To date, the Institute of Education Sciences (IES) has awarded 41 states and the District of Columbia over \$514,000,000 in four rounds of grants for states to create and expand state longitudinal data systems (SLDS).(The Institute of Education Sciences [IES], 2011) Beginning with the fourth round of grants, and extending to the fifth round which has not yet been awarded, the IES extended the focus of these grants to expand beyond K12 and encourage states to create P20W systems, which links data on individuals from pre-Kindergarten, through elementary and secondary education, higher education, and on to the workforce. Of the 20 states awarded funding in the fourth round of SLDS grants, 18 are creating data systems that link at least two agencies, 8 of which are creating P20W systems.("State Longitudinal Data Systems; Grantee States," n.d.)

There were already concerns about individual privacy with the creation of K12 SLDS programs. Joel Reidenberg, who oversaw a study by Fordham's School of Law Center on Law and Information Policy, said that with the creation of such systems states "are trampling the privacy interests of those students."(Anderson, 2009, para. 3) Extending these systems to include data from agencies outside K12 and including agencies that are not education providers prompted the U.S. Department of Education to make changes to FERPA regulations to help facilitate the creation of such P20W systems. The ACLU expressed its concern with these changes, arguing not only that it allows for the non-education providers to view student data but that it also could lead to the creation of a national database of individual student data:

Personally identifiable student records include extremely sensitive information about individuals, yet these rules significantly expand the number of parties who can access a record without requiring consent from the parent or the student. These new parties include state officials not working directly on education as well as private entities that would not traditionally be able to access government educational records. Furthermore, the expansion of access to student records could eventually lead to sharing among states. If this were to happen, it could lead to the creation of an immense database holding sensitive information about most Americans. (Murphy & Calabrese, 2011, p. 6)

Besides the privacy concerns over the creation of such data systems, there are technical issues as well. Data definitions are unlikely to be consistent across agencies nor are data attributes. The federal government has initiated the National Information Exchange Model to help with such issues, and similar work in the education domain in support of P20W systems has been taking place with the Common Education Data Standards (CEDS). Such efforts are demonstrative of the drive to link disparate agency data at the federal, state, and local level. While it may seem a stretch to construe that such efforts exist at the federal level, the U.S. Department of Education's efforts in the area of "Gainful Employment" are but one example that require the merging of individual-level data at the federal level.("Gainful Employment," 2010)

One of the most challenging problems facing such implementations would be identity management. How can we be sure that one record from Agency A is tied to the same person as a record in Agency B? SSN is most commonly used to link such data together, but a link on SSN alone is not as strong as might be imagined. In Arkansas, we have found 38,137 SSNs that are shared by two or more individuals. That means if we were to use SSN alone for matching, which is quite common in other states, we would have a conflict with 1 in 83 people. To further complicate the problem, parents are allowed to enroll a student without providing their child's actual SSN, so such students are entered in the K12 system with a SSN that begins with a 9, which is not valid under existing SSN creation rules. There are currently 14,057 students with such an ID, but this number has been as high as 20,000 in previous years. Records with such an ID will not be able to be matched with records from agencies that do require a valid SSN. The fact that Arkansas has multiple individuals "sharing" a single SSN and that some individuals exercise their right not to include their actual SSN as part of a data collection are both merely demonstrative that a problem exists and provides little insight into the actual scope of potential SSN problems. Again, it is important to note that matching via SSN alone is the primary means by which most states are currently creating systems the combine data from multiple agencies.

To ensure the greatest number of possible entity matches, multiple attributes of PII are needed for entity resolution, but again, data quality issues and the non-uniqueness of some data elements make this difficult. Of the 38,137 entities above that share their SSN with at least one other individual, 2,730 of them also share the same data of birth. (G. Holland, personal communication, December 10, 2011) Approximately 29% of Arkansans share the same first and last name.(G. Holland, personal communication, December 10, 2011) Name, date of birth, and

SSN are all easily susceptible to such things as transposition of characters, and names can change with new life experiences such as marriage, divorce, or adoption.

To deal with these issues, the Arkansas Research Center (ARC) created an open-source Knowledgebase Identity Management program (KIM), which it has shared with other states. KIM maintains all representations of an entity in a master index, in order to facilitate matching. If a record for Kathy Jones is matched using first name, SSN, and date of birth to Katherine Smith, both “Jones” and “Smith” are maintained in the system to provide a greater chance of matching a third record that may have either of those last names, even if the date of birth or SSN might be different because of transposed characters or other problems. A Knowledgebase ID is generated to identify clusters of records that belong to the same entity. KIM uses a stepwise process of both exact and approximate matching that does not rely exclusively on SSN alone and that can be easily modified to a user’s particular needs or the requirements for a particular data set. The program can be downloaded from the Arkansas Research Center website. (Arkansas Research Center website, n.d.)

KIM represents one half of the TrustEd framework. Besides entity resolution, KIM also generates an agency specific ID which is an encryption of KIM’s Knowledgebase ID concatenated with an agency identifier. All PII is maintained within the KIM system which does not contain any data of research interest. The Agency ID is appended to the research data, from which all PII is removed. This research data, with the appended Agency ID, is then loaded into an agency specific database or “edge server.” Under this arrangement, there are two levels of privacy protection. The data exists in a completely de-identified manner without any PII, and there are no direct links between one agency’s data and any other agency’s data.

If a research request requires the integration of data between two or more agencies, a crosswalk of Agency IDs is generated via KIM, and this crosswalk is then loaded into a new database instance. Using this crosswalk as a bridge between the edge servers, the required dataset can be constructed. Datasets are not allowed to be constructed that include any Agency IDs. One of the Agency IDs is normally encrypted for the resulting data. In other cases a new, specific Research ID is created for partners that request data be updated on a regular basis, so they can build their own longitudinal systems with these de-identified data. Once the result set is generated, the temporary crosswalk is destroyed and there are no longer any links between agency data. It is also important to note that if necessary, there is a means by which, although very difficult and resource intensive, a particular individual that is part of a research study can be eventually determined if needed. Such processes are required under certain research protocols, such as those involving the National Institutes of Health.

While this may seem somewhat of a simplistic approach, it does solve many issues related to the creation of state longitudinal multi-agency data systems. Creators of such systems have to be mindful of the “Big Brother” nature of their work and why such work raises concerns among a variety of stakeholders. It is not enough that such data be de-identified; we must also ensure that such data cannot be easily subjected to data mining, where linked data, even if it is de-identified, could be indiscriminately subjected to the process of automated pattern recognition. In the case of TrustEd, any potential query of data across multiple agencies must begin with a specific research question in mind. Only after the research inquiry has been vetted and the protocols agreed to by all agencies involved is the necessary crosswalk created and the needed dataset produced. By keeping PII data in a system separate from the system that holds the research data of interest, and by keeping the various agency data de-identified at the agency

level, data maintained by ARC is at a level of anonymity and protection of PII unmatched by any other state that currently has such a multi-agency system.

The current state of TrustEd allows the state of Arkansas to do cutting-edge research while maintaining the protection of PII and preventing the exposure of such data to arbitrary machine learning algorithms beyond the traditional scope of empirical research. An example of such research would be the wage outcomes of college graduates by post-secondary certifications and Classification of Instructional Programs codes. (Walker & Holland, 2011) ARC is currently extending this research to include information about those which have only a high school diploma, General Educational Development diploma, high school dropouts, and high school graduates with some college hours. Such research also extends to the areas of factors determining college success, the impact of Advanced Placement courses on college success, how early learning programs lead to improved academic achievement of disadvantaged children, and the educational outcomes of infants born with a variety of medical conditions up to twelve years after their birth.

Such research topics would be the envy of any state creating a similar longitudinal data system, and Arkansas is very proud that it has been able to create such a system that can answer these questions while keeping individual privacy at the center of this system's creation. However, this is not the end state we envision for TrustEd. Our eventual goal is to create a framework in which individual agencies maintain their own edge servers, and the population of Agency IDs as well as the querying of multiple agency data is all done via web-services protocols. While we are not yet at this level of ability, we are satisfied that our current state of capacity is well beyond that of but a handful of states, and we are equally proud that we were

able to reach this state while maintaining a level of individual privacy and protection unmatched by any other state currently creating a similar system.

References

- Anderson, N. (2009, October 28). States not protecting student privacy, study finds. *The Washington Post*. Retrieved from <http://www.washingtonpost.com/wp-dyn/content/article/2009/10/27/AR2009102703562.html>
- Arkansas Research Center website. (n.d.). <http://arc.arkansas.gov>
- Department on Track to Implement Gainful Employment Regulations. (2010). Retrieved from <http://www.ed.gov/news/press-releases/department-track-implement-gainful-employment-regulations-new-schedule-provides->
- Grantee States. (n.d.). Retrieved from <http://nces.ed.gov/programs/slds/stateinfo.asp>
- Murphy, L. W., & Calabrese, C. R. (2011, May 23). *Re: Family Educational Rights and Privacy Act (FERPA) regulatory changes, Docket ID ED-2011-OM-0002* [Response Letter]. Retrieved from ACLU website: <http://www.aclu.org>
- The Institute of Education Sciences. (2011). UPDATE: SLDS Grant Competitions. Retrieved from <http://ies.ed.gov/whatsnew/newsletters/nov11.asp?index=gng>
- Walker, J., & Holland, G. (2011). *Arkansa education to employment report 2011* . Retrieved from Arkansas Research Center website: http://arc.arkansas.gov/arc/resources/AEER_FullReport_FINAL_20120113.pdf