SHEEO

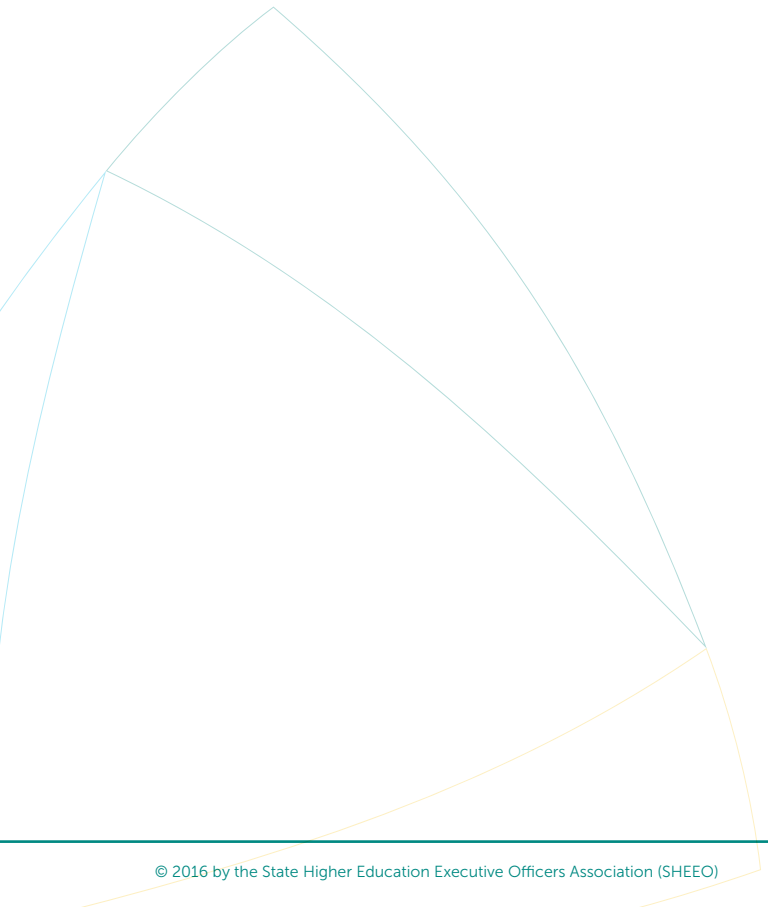
STATE HIGHER EDUCATION EXECUTIVE OFFICERS ASSOCIATION

PRIVACY, CONFIDENTIALITY, AND SECURITY IN ARKANSAS

EFFECTIVE USE OF STATE DATA SYSTEMS

CHRISTINA WHITFIELD

This paper is based on research funded in part by the Bill & Melinda Gates Foundation. The findings and conclusions contained within are those of the author(s) and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation.



Analysis of student-level data to inform policy and promote student success is a core function of executive higher education agencies. Postsecondary data systems have expanded their collection of data elements for use by policymakers, institutional staff, and the general public. State coordinating and governing boards use these data systems for strategic planning, to allocate funding, establish performance metrics, evaluate academic programs, and inform students and their families. The State Higher Education Executive Officers Association (SHEEO), as part of a project funded by the Bill & Melinda Gates Foundation (BMGF), surveyed state coordinating and governing boards on their collection and use of postsecondary student-level data. Following this, SHEEO identified seven states whose survey responses indicated an exemplary use of data in specific subject areas. In-person interviews were conducted by SHEEO agency staff in seven states selected for follow-up. In 2015, SHEEO visited the Arkansas Research Center to discuss their efforts to promote privacy, confidentiality, and data security.

Postsecondary data practitioners face dual calls to action regarding the use of postsecondary student unit records systems (PSURSs). On one hand, policymakers call for increased use of data and information contained in these systems to improve higher education outcomes. On the other hand, the public has heightened concerns regarding big data and data breaches. An example of a PSURS that effectively balances data use and information accessibility with protecting privacy and maintaining security may be found at the Arkansas Research Center (ARC). The staff at ARC—which functions as a state longitudinal data system (SLDS) in the state—are very cognizant of the broader context regarding privacy and security in which they operate. They have developed a sophisticated approach to safeguarding information, which includes a philosophical understanding of privacy and confidentiality and a multi-layered technical approach to data security.

Over the past decade, a broad consensus has emerged among researchers and policymakers that more and better data are needed to inform postsecondary education's various constituencies and improve performance and student success. In 2014, the Institute for Higher Education Policy (IHEP) issued a report outlining significant gaps in the types of vital data and information needed to inform students, policymakers, and postsecondary institutions.[1] In 2016, the BMGF, building on the work of numerous voluntary data collection and accountability initiatives, developed a comprehensive framework of metrics designed to "provide the information necessary to improve the capacity and productivity of the higher education system," while acknowledging that existing postsecondary data systems are insufficient to fulfill the framework's vision (more and better data will be necessary to implement the framework).[2]

In response to these calls for better data and greater access to information, and to changes at the state level, including the proliferation of performance funding models, PSURSs are becoming more complex and are increasingly linked to other data systems. SHEEO's "The State of State Postsecondary Data Systems" report outlines the wide range of data elements housed within PSURSs, and the dramatic growth in the number of postsecondary systems that link with K-12, workforce, and other data systems.[3]

1. Mamie Voight, Alegneta A. Long, Mark Huelsman, and Jennifer Engle, "Mapping the Postsecondary Data Domain: Problems and Possibilities," Institute for Higher Education Policy (IHEP), March 2014.
2. Jennifer Engle, "Answering the Call: Institutions and States Lead the Way Toward Better Measures of Postsecondary Performance," Bill & Melinda Gates Foundation, 2016: 22.
3. John Armstrong and Christina Whitfield, "The State of State Postsecondary Data Systems: Strong Foundations 2016," SHEEO, May 2016.

Along with the growing use and complexity of PSURSs comes widespread concern about privacy, confidentiality, and security. As the Data Quality Campaign indicates, "questions from the public about how these data systems work and how student privacy is protected have been increasing," and "student data privacy has emerged as a prominent theme in policy, media, and political conversations."[4] Anxiety about data privacy and security is fueled by well-publicized data breaches outside the postsecondary data sphere. The exposure of millions of records held by corporations and government entities, such as those at Target, Anthem, and the US Office of Personnel Management, has increased public sensitivity to the vulnerability of "big data." The rise and fall of InBloom, a non-profit formed to promote personalized learning and improve outcomes for K-12 students, has been attributed to "the company's failure to convince people it adequately protected the data."[5]

This context of public concern and the challenges inherent in maintaining privacy and security are explicitly addressed by leaders in the higher education data policy realm. As Archie Cubarrubia and Patrick Perry state in the introduction to a recent series of papers on improving the national postsecondary data landscape: "Creation of an agile and effective postsecondary data ecosystem cannot come at the expense of the privacy of the individuals whose personal and educational records are contained within it."[6] As Joanna Lyn Grama writes in a paper within that series, "honoring the privacy of students and families represented in these [postsecondary data] systems—while also using the data to inform decisions and improve outcomes—is an effortful endeavor."[7]

It is widely acknowledged that "the regulatory environment affecting postsecondary data collection and storage activities is complex."[8] This regulatory environment has, in the past, been dominated by concerns about the Family Educational Right to Privacy Act (FERPA)—the basis for federal requirements around a student's right to privacy at all levels, including postsecondary education. Although FERPA was formerly perceived as a significant barrier to data sharing and analysis, a consensus has emerged that data linkages and educational research can be undertaken successfully under the auspices of FERPA.[9] While FERPA serves as the fundamental underpinning of federal efforts to protect student data, there are numerous other federal laws that protect data elements that might be found in PSURSs.[10] PSURSs practitioners must also be aware of the evolving state legislative landscape. Responding primarily to parental concerns about protecting children's data privacy while engaged with K-12 systems, legislators in many states have introduced laws aimed at protecting data privacy.[11] In 2014, 36 states introduced 110 bills on student privacy; in 2015, 46 introduced 182 bills, 28 of which (including one in Arkansas) became law.[12] The impetus for much of this legislation is positive—protecting minor students' personally identifiable information (PII). However, the impact of these bills can have an inadvertent chilling effect on postsecondary

4. Data Quality Campaign, "Student Data Collection, Access, and Storage: Separating Fact from Fiction," October 2014. Retrieved from: *http://dataqualitycampaign.org/resource/199*
5. Olga Kharif, "Privacy Fears Over Student Data Tracking Lead to InBloom's Shutdown," Bloomberg, May 2, 2014. Retrieved from: *http://www.bloomberg.com/news/articles/2014-05-01/inbloom-shuts-down-amid-privacy-fears-over-student-data-tracking*
6. Archie Cubarrubia and Patrick Perry, "Creating a Thriving Postsecondary Education Data Ecosystem," IHEP, May 2016, 4.
7. Joanna Lyn Grama, "Understanding Information Security and Privacy in Postsecondary Education Data Systems," IHEP, May 2016, 2.
8. Ibid.
9. See Armstrong and Whitfield, 26-7.
10. For a guide to these laws, see Grama, 4.
11. Andrew Ujifusa, "State Lawmakers Ramp Up Attention to Data Privacy," *Education Week*, April 15, 2014. Retrieved from *http://www.edweek.org/ew/articles/2014/04/16/28data.h33.html*
12. Data Quality Campaign, "Student Data Privacy Legislation: What Happened in 2015, and What is Next?" September 2015. Retrieved from: *http://dataqualitycampaign.org/resource/student-data-privacy-legislation-happened-2015-next*. The Arkansas law prohibits service providers from using student data to target advertising or disclosing student information. See Tanya Roscorla, "More States Pass Laws to Protect Student Data: Legislatures in 15 states passed 28 laws to safeguard the privacy of student data this year," Center for Digital Education, August 27, 2015. Retrieved from: *http://www.centerdigitaled.com/k-12/What-States-Did-with-Student-Data-Privacy-Legislation-in-2015.html*

education research.[13] In a context in which increasing numbers of PSURSs link with K-12 data, and to the extent this legislation informs the general atmosphere regarding data privacy and security, these bills cannot be ignored by postsecondary practitioners.[14]

In the context of this challenging environment regarding privacy, confidentiality, and security, the Arkansas Research Center functions as an exemplar for other PSURSs. ARC was established in 2009 via funding from the Institute for Education Statistics (IES). Its mission is to "provide educators, parents, policymakers, and researchers with relevant data to improve educational outcomes for students in Arkansas."[15] ARC promotes data-based decision-making and data visualization, and links (or linked) data across multiple state agencies (including the Arkansas Department of Human Services, Arkansas Head Start, the University of Arkansas for Medical Sciences, and the Arkansas Departments of Education, Higher Education, Workforce Services, and Corrections) to facilitate education- and workforce-related research.

ARC was established under the auspices of the Arkansas Commission for the Coordination of Educational Efforts (ACCEE). ACCEE, formed in 2003 by the Arkansas legislature, makes recommendations on P-16 policy, and encourages cooperation among K-12 education, higher education, and the workforce.[16] ARC currently operates as a part of the University of Central Arkansas (UCA), which is the signatory for all ARC contracts. ARC, like other SLDSs, is required to seek institutional review board (IRB) approval for inter-agency research projects. UCA's IRB provides oversight for ARC projects.

ARC staff emphasize the following themes in their work:

- Development and improvement of TrustEd, their in-house "trusted broker data model"
- Data visualization and promotion of ease of access to data
- Cross-agency data sharing and coordination of research activities among state agencies
- Data-based decision-making and professional development around appropriate use of data
- Development and promotion of national data standards[17]

During the early period of ARC's existence, these themes were most often reflected in the Center's work with the state's K-12 system and in linking postsecondary education and workforce data. More recently, the bulk of the Center's work has shifted to support for Arkansas's workforce agencies.

13. John Armstrong and Katie Zaback, "Assessing and Improving State Postsecondary Data Systems," IHEP, May 2016, 13.
14. The complex and evolving regulatory landscape described here largely concerns what might be deemed "traditional" educational data—data found in static files in state longitudinal data systems or transactional data in institutional systems. A set of privacy concerns and ethical challenges is emerging around the proliferation of learning and early warning systems and predictive analytics. These platforms yield a wealth of "clickstream data," detailed information about student behavior and online activities. The availability of these data means that institutions increasingly have the capacity to predict the likelihood of student success in particular courses or programs. Questions about whether it is appropriate to do so, or what an institution's obligations are regarding these determinations, has led to calls for "ethical standards around educational data that go beyond legal issues of privacy or security." See Goldie Blumenstyk, "As Big Data Comes to College, Officials Wrestle to Set New Ethical Norms," *The Chronicle of Higher Education*, June 28, 2016. Retrieved from: *http://www.chronicle.com/article/As-Big-Data-Comes-to-College/236934*
15. *https://arc.arkansas.gov*
16. See *https://arc.arkansas.gov/governance/ACCEE* and *http://ecs.force.com/mbdata/MBProfSN?SID=a0i700000009vZI&Rep=PSST&state=Arkansas*
17. *https://arc.arkansas.gov/what*

The seriousness of ARC's commitment to privacy and the innovative nature of their approach to security are widely acknowledged. Dr. Neal Gibson (director of the Center from 2013 to 2016) and Dr. Greg Holland (current director) frequently write about, and present on, their privacy and security practices.[18] The IES, in an *SLDS Spotlight report*, notes that state recipients of SLDS grants are "tasked with protecting the anonymity and privacy" of students whose data is housed within these systems, and that the "Arkansas Research Center has taken unprecedented strides to this end."[19]

Though widely known for the technical aspects of their privacy work, ARC's treatment of privacy and security begins with a philosophical, rather than a technical, stance. Gibson grounds his definitions of privacy and confidentiality (differentiating both of these from security) in three somewhat surprising sources, none of which are primarily concerned with either data or postsecondary education:

- First is "The Right to Privacy," an 1890 essay by Samuel Warren and Louis Brandeis. During the interview with SHEEO staff, Gibson remarked that any discussion of privacy should begin with a definition of the term. "Believe it or not, it's kind of a new concept . . . I think the best definition is from Louis Brandeis . . . He was the one that really pushed this to the consciousness of American thought." In their essay, Warren and Brandeis outline the evolution of the concept of personal rights, and define privacy as the "right to be let alone." Like those concerned about the vulnerabilities of "big data" today, Warren and Brandeis were responding to what they perceived as an emerging threat to privacy, in their case, to journalistic practices that included gossip columns and tabloid photography.[20] Warren and Brandeis "conceived of the right to privacy as a kind of presumption of individual control over personal information,"[21] and it is this presumption that provides a connection between late nineteenth century concerns and modern concerns about data tracking within PSURSs.
- The second influence mentioned by Gibson is Kenneth Prewitt, former director of the US Census Bureau. In a 2011 article, Prewitt argues that policymakers and researchers cannot fully address public concerns about privacy without differentiating between privacy and confidentiality. "At the most simple and most common-sense level," Pruitt writes, "the distinction is between *'don't ask'* and *'don't tell.'*"[22] In the interview, Gibson noted "you can protect confidentiality with security, but you're not protecting, necessarily, privacy." In a recent essay, Gibson and Holland explain that "the issue of privacy, 'don't ask,' is central to discussions concerning the joining of data between different agencies." While individuals

18. For examples, see:

    Neal Gibson, "Data Privacy and Confidentiality" presentation, SHEEO Meeting on Effective Utilization of Postsecondary Data Systems, Boulder, CO, December 8, 2015. Accessible at: *http://www.sheeo.org/sheeo-meeting-effective-utilization-postsecondary-data-systems*;

    Regional Education Laboratory Northeast and Islands, "We Built Longitudinal Data Systems; Can't We Find Time to Use Them?" May 19, 2014. Retrieved from: *http://www.relnei.org/news/longitudinal-data-systems-event-summary.html*;

    Neal Gibson and Greg Holland, "A Dual-Database Trusted Broker System for Resolving, Protecting, and Utilizing Multi-Sourced Data," in *Information Quality and Governance for Business Intelligence* by William Yeoh, John R. Talburt, and Yinle Zhou, IGI Global, 2014.

19. IES, "SLDS Spotlight: Arkansas's Approaches to Identity Management," April 2013. Retrieved from *https://nces.ed.gov/programs/slds/pdf/AR_spotlight.pdf*

20. Samuel Warren and Louis Brandeis, "The Right to Privacy," *Harvard Law Review* Vol. IV, No. 5, December 15, 1890. Retrieved from: *http://groups.csail.mit.edu/mac/classes/6.805/articles/privacy/Privacy_brand_warr2.html*

21. Dorothy Glancy, "The Invention of the Right to Privacy," *Arizona Law Review* Vol. 21, No. 1, 1979. Retrieved from: *http://digitalcommons.law.scu.edu/cgi/viewcontent.cgi?article=1318&context=facpubs*

22. Kenneth Prewitt, "Why It Matters to Distinguish Between Privacy & Confidentiality," *Journal of Privacy and Confidentiality*, Vol. 3, No. 2, 41-47, 2011. Retrieved from: *http://repository.cmu.edu/cgi/viewcontent.cgi?article=1073&context=jpc*

might willingly divulge information about themselves to an agency in order to receive services, they may do so with an expectation of privacy—an expectation that that agency will not divulge PII to others.[23]

- Finally, Gibson referenced "The Belmont Report," or the "Common Rule." The existence of complex, linked data systems means that researchers may pose myriad research questions. Yet the existence of this capacity does not mean that all research questions are appropriate. For Gibson, an important aspect of confidentiality protections is determining whether "this an okay question to ask." "The Belmont Report," issued by the US Department of Health & Human Services in the 1970s, lays out the ethical principles for research on human subjects, and serves as the basis of decision-making for institutional review boards in the United States.[24] Notes Gibson: "We are at the cusp of this convergence where we're going to able to ask all kinds of things. There has to be something in place that asks whether or not that's allowable." Gibson's desire for this guidance led to his early insistence on the oversight of an institutional review board for ARC.

ARC's philosophical approach to privacy and confidentiality is supported by sophisticated data security methods. An important aspect of this is their operating assumption that the security of their data is at risk. ARC's vigilance regarding security is fueled by an awareness of potential threats from both outside and inside the system. Matt Jeffery (ARC's lead software developer) remarks of external threats: "One of the assumptions that we operate under here is that we're going to get hacked." Regarding internal threats, Gibson notes that what was "arguably the biggest data breach of all time [Edward Snowden's leaking of National Security Agency information] was not a hack . . . [It was] an inside job." Because no system can be secure enough to protect against every possible hacker or rogue employee, ARC's database is designed so that if compromised, no PII will be found. TrustEd, ARC's multi-layered approach to privacy and security, is predicated on the idea that there is no single, infallible way to protect data. Rather, the system employs "defense depth" security, multiple layers of protections that, taken together, provide a high likelihood of protecting the information in the system even if individual elements are less than perfect.

For Jeffery, the goal of ARC's database design is to "do the hard stuff"—make PII inaccessible—while "making it easy for researchers and users to have access to what's actually useful to them." ARC's complex technology protects privacy and security, while allowing researchers and policymakers access to data from multiple agencies. For Holland, providing ease of access to researchers in a context of highly secured data is the ultimate "win-win" situation. ARC does this in part through its federated data model. Rather than storing data from multiple agencies together in a single "warehouse," data from various sources are stored separately and joined only for express purposes allowed by their data governance structure and to answer specific research questions. ARC sees this federated model as an additional aspect of privacy protection. Storing the data separately prevents individuals with access to the data from playing "Go Fish." "The right to individual privacy," Gibson and Holland write, "trumps any need . . . a researcher might have to 'explore' the data."[25]

ARC's system for managing data and protecting privacy, confidentiality, and security is its dual database architecture, TrustEd. TrustEd is comprised of three modules: Knowledgebase Identity Management (KIM), TrustEd Identifier Management (TIM), and agency-specific de-identified research databases.[26]

23. Gibson and Holland, 354.
24. *The Belmont Report*, Washington, D.C. Government Print. Office, 1978, was inspired by a series of unethical medical experiments conducted on African-American men in Alabama in the 1940s. Retrieved from: *http://www.hhs.gov/ohrp/regulations-and-policy/belmont-report*.
25. Gibson and Holland, 354.
26. For more detailed descriptions of these modules and processes, see: *https://arc.arkansas.gov/trustEd*, and IES, "SLDS Spotlight: Arkansas's Approaches to Identity Management."

- KIM is the module within TrustEd that performs identity management. KIM uses PII to determine whether new information entering the system should be tied to an existing record or represents a new individual (e.g., are "Kate Smith," "Katie Smith," and "Katherine Smith" the same person, or several people?). Many cross-agency data matching systems rely on Social Security numbers to perform these kinds of matches. KIM uses a number of statistical models and additional data elements to accomplish a higher and more reliable match rate than reliance on SSN alone. KIM is the only module within TrustEd that contains PII, or, as Gibson referred to it in the interview, "the plutonium." One of the added layers of security within TrustEd is that the KIM database (which contains SSN, name, and date of birth) is used only rarely—when a new file is received from an agency and identities need to be resolved. For most of the time, the server that houses this database is powered down (literally turned off), providing a physical barrier to accessing PII.

- TIM is the module within TrustEd that constructs temporary crosswalks to link data across agencies. TIM creates a temporary identification and crosswalk specific to a research request, which is subsequently destroyed. For example, if an approved research request is received regarding exploring employment outcomes for college graduates, TIM will temporarily link records within multiple databases. The system will determine that the de-identified record "37" (no PII is involved) in the postsecondary education database and the de-identified record "943" in the wage records database are the same person. TIM creates a temporary code to link these records and makes the linked data available to the researcher. Once the research request is fulfilled, the link between record 37 and record 943 is destroyed.

- De-identified research databases are generated by the process described above, and made available to the requesting agency. ARC functions as the "trusted broker" in these circumstances, and data governance plays a key role in the process. Continuing with the example cited above, ARC performs links between postsecondary education and employment data only if the appropriate agencies (i.e., the Arkansas Department of Higher Education and the Arkansas Department of Workforce Services) agree to the appropriateness of the project.[27]

Whether ARC's approach is entirely replicable in other states is an important question. TrustEd is a homegrown solution; it does not rely on any vendor software. This approach is advantageous in that it avoids costs associated with third-party vendors, but, because it relies on the advanced technical expertise of ARC staff, is potentially difficult to replicate. The staff at ARC have expressed a willingness to share their work, and have had conversations with several states about implementing versions of TrustEd elsewhere. While several states have adopted the TrustEd model (or some version of the model), no other state has adopted the TrustEd software. This speaks to the strengths and challenges of what ARC built. As the TrustEd website states: "We realize that this approach is both more complex and burdensome than perhaps other longitudinal data systems. However, we believe this system provides some important capacities and extra privacy protections."[28]

27. ARC plans to add yet another layer of security to this data by encrypting the information within these databases. Currently, Katherine Smith's information is encoded as a series of random digits. In the future, ARC intends to embed her information within a series of encoded strings that, without the decryption key, cannot be deciphered. Gibson, "Data Privacy and Confidentiality."

28. *https://arc.arkansas.gov/trustEd*. ARC benefits from the industry experience of two of its key employees. Both Holland and Jeffery came to ARC from Acxiom, a for-profit enterprise in Conway, Arkansas, that specializes in consumer information, and has developed advanced mechanisms for matching identities across multiple databases while protecting personal information.

ARC's approach to privacy, confidentiality, and data security provides a gold standard for PSURSs. While certain technical aspects of their approach may not be easily replicable, other features of ARC's approach can inform a wide range of systems. Recommendations for other states based on ARC's system include:

- Seriously consider how your system protects privacy, confidentiality and security.
- Assume your system is subject to external and internal security threats.
- Design a multi-layered system that protects student privacy.
- Use a data governance structure to determine the appropriateness of inquiries.
- Ensure ease of access for researchers and policymakers with legitimate inquiries

STATE HIGHER EDUCATION EXECUTIVE OFFICERS
3035 CENTER GREEN DRIVE, SUITE 100, BOULDER, COLORADO, 80301
303.541.1600 • **SHEEO.org**