Understanding Information Security and Privacy in Postsecondary Education Data Systems

JOANNA LYN GRAMA
EDUCAUSE

MAY 2016

EDUCAUSE

POSTSEC DATA

Joanna Lyn Grama, **JD**, **CISSP**, **CIPT**, **CRISC**, directs the EDUCAUSE Cybersecurity Initiative and the IT GRC (governance, risk, and compliance) program. She is a frequent speaker on a variety of IT security topics, including identity theft, personal information security, and university information security compliance issues. She is also the author of the textbook, *Legal Issues in Information Security* (2nd Ed., July 2014).

This paper is part of the larger series *Envisioning the National Postsecondary Data Infrastructure in the 21st Century.* In August 2015, the Institute for Higher Education Policy (IHEP) first convened a working group of national postsecondary data experts to discuss ways to move forward a set of emerging options for improving the quality of the data infrastructure in order to inform state and federal policy conversations. The resulting paper series presents targeted recommendations, with explicit attention to related technical, resource, and policy considerations. 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 or the Institute for Higher Education Policy.

Executive Summary

The current national postsecondary education data infrastructure is insufficient to provide decision makers with the data they need most about students and their outcomes. Better data, acquired through both existing initiatives and through other options currently under consideration by education stakeholders, are required to provide meaningful information about student outcomes and to improve higher education. Although stakeholders have expressed concerns about student privacy and information security in light of the growing need for better data, it is possible to protect students and sensitive information while also developing effective data collection strategies and systems.

With thoughtful planning, comprehensive information security and privacy practices can be implemented within the national postsecondary education data ecosystem. For this planning to be successful, all stakeholders in the ecosystem must have a foundational understanding of basic information security and privacy concepts. They also must recognize the most important "big data" information security and privacy concerns such as volume, sensitivity, and access. Finally, while there is no exact formula to guarantee information security and privacy with any data or information technology (IT) solution, stakeholders in the ecosystem should adopt a risk-based approach to thoroughly protecting data in the education ecosystem.

Prior research has indicated that data collection activities designed to solve issues of equity and meaningfully contribute to positive student outcomes are necessary.[1] These data must be protected after they are properly collected. This paper outlines the information security and privacy challenges that exist within the national postsecondary education data infrastructure. It also introduces key language related to information security and privacy in order to provide a common lexicon to frame data protection discussions, describes the general technology architectures considered in the national postsecondary education data infrastructure, summarizes big data information security and privacy concerns, and outlines information security and privacy best practices that could be used to protect data within the national postsecondary education data ecosystem.

Discussions about information security and privacy often fail to meet desired outcomes because technologists and policymakers are not using the same language to describe data protection outcomes. Often concepts are "lost in translation" when both parties do not understand one another. This is quite common when policy and regulatory concepts must be reduced to technological controls that then must be applied to IT systems. To make sure that this does not happen in conversations around and in the further development of the national postsecondary education data ecosystem, it is incumbent upon all stakeholders in the system to understand information security and privacy concepts as they relate to big data.

The current national postsecondary education infrastructure is complex and has many stakeholders and many underlying IT systems. State and/or federal government action likely will be required to successfully steward student data, which includes robust information security and privacy practices. Information security and privacy must be a foundational element of any national postsecondary data system. With intentional, collaborative planning, stakeholders within the national postsecondary education data infrastructure can implement the necessary information security and privacy practices that reduce risk, safeguard data, and ensure transparency, accountability, and trust throughout the entire ecosystem. The following four recommendations form a holistic framework for ensuring effective information security and privacy protections within the national postsecondary education data ecosystem:

1. Adopt a risk-based approach to understanding information security and privacy threats and vulnerabilities.

2. Establish and adhere to a baseline set of information security protections.

3. Establish and adhere to a baseline set of privacy standards.

4. Implement a collaborative governance structure that includes addressing information security and privacy throughout the national postsecondary education data infrastructure.

Understanding Information Security and Privacy in Postsecondary Education Data Systems

Introduction

Mapping the Postsecondary Data Domain: Problems and Possibilities, a March 2014 Institute for Higher Education Policy (IHEP) report, has outlined the critical questions higher education stakeholders must ask and the core measures they must implement in order to provide students, policymakers, and institutions useful and reliable information about student success in the postsecondary education system.[2] Better data, acquired through both existing initiatives and other options currently under consideration, are required to provide meaningful information about student outcomes in order to improve higher education.[3] Researchers have widely acknowledged that the data currently available are inadequate and that change is required for improved data collection, sharing, and analyses within the national postsecondary education data infrastructure.[4]

Data provide the infrastructure of the information age. Every postsecondary education institution in the United States collects and uses data in their enterprise-level information technology (IT) systems. State, regional, and federal entities also collect and use data. Data, and all the IT systems in which they are maintained, are high-value strategic assets and must be respected and protected. Securing these assets and honoring the privacy of students and families represented in these systems—while also using the data to inform decisions and improve outcomes—is an effortful endeavor. All stakeholders, from students and parents to administrators and policymakers, must have confidence that students' personally identifiable information is collected only when necessary and lawful, used appropriately, and is reliable. The regulatory environment affecting postsecondary data collection and storage activities is complex; stakeholders in the system may have differing privacy expectations, and security safeguards must be carefully crafted for all IT systems in the ecosystem. Balancing the opportunities postsecondary education data systems present in terms of understanding the factors that contribute to improving student success with the risks to information security and privacy is more art than science.

Understanding how information security and privacy concepts work together to protect data across the national ecosystem[5] of postsecondary education data is crucial. The data that must be analyzed to feed the required measures and answer the critical questions posed in *Mapping* may come from a variety of data owners, including students, institutions, non-governmental agencies, and state and federal actors; may reside in a number of IT systems controlled by these different entities; and may require varying levels of sensitivity (e.g., identifiable student-level data versus de-identified and aggregate data). Some of the data collected may be protected by federal and/or state law. Other data might not be protected by law but could be highly sensitive to the associated individual and embarrassing if shared or disclosed. Understanding the complexities of how data are collected and the legal issues at play is necessary to make informed decisions for policy and practice. Since this ecosystem is so complex, it is likely that state and/or federal government action will be required to successfully steward student data throughout the ecosystem and to produce the meaningful collaboration that provides decision makers with the data they most need to understand student outcomes.

This paper seeks to inform the reader of the information security and privacy concepts and challenges that must be considered and addressed in the national postsecondary education data infrastructure, whatever its configuration. To that end, this paper provides the following:

- A foundational understanding of basic information security and privacy concepts and the language that technologists use to explain these concepts
- A description of general technology infrastructure architectures considered in the national postsecondary education data infrastructure
- A summary of the information security and privacy concerns inherent in big data collections such as those that are part of the national postsecondary education data infrastructure
- An outline of information security and privacy concepts to be considered in any national postsecondary education data infrastructure solution(s)

Information Security and Privacy Concepts: A Common Understanding

Data are the core assets of any information technology system designed to provide users with reliable information on which to base their decisions. Student-level data (e.g., admission, matriculation, progress, completion, and outcome data)

are essential to providing a meaningful national data solution, and information security and privacy protections must be a foundational element of any system in order to protect these data. All stakeholders in the ecosystem must have a shared understanding of *information security* and *privacy* that extends beyond a lay interpretation of these concepts, including how these concepts work together to protect student data and ensure that data are used properly within the institutional, state, regional, and federal IT systems that compose the overall ecosystem infrastructure.

What Is Information Security?

Information security refers to the mechanisms that protect data. Often those less familiar with information security consider it a mere technical control implemented into IT systems. In reality, however, information security is more than a mere technical control and must be understood as the study and practice of protecting data in all its forms (e.g., whether stored in an IT system or reduced to paper or another physical medium). It includes protecting data from all types of threats, whether those threats are perpetrated by malicious outsiders or individuals with legitimate accesses to IT systems and data. The practice of protecting data includes three distinct information security concepts, outlined in the following paragraphs.

Confidentiality means protecting data, in all its forms, from unauthorized access throughout its entire lifecycle (from data creation to data destruction). Unauthorized access includes access by individuals not affiliated with the underlying organization storing the data (e.g., criminals and hackers). It also includes access by individuals within an organization who purposefully exceed their scope of authority in accessing information (e.g., individuals looking up the records of celebrities or other targeted individuals when they have no professionally legitimate reason to do so).[6] Confidentiality is the information security concept most often implicated when an organization experiences a data breach.[7]

Integrity means ensuring that data within IT systems (or recorded or reproduced on physical media) are accurate. This means that IT system creators and managers implement controls within the system to ensure that users enter and process data correctly and that conflicting data elements are identified and resolved. Integrity also requires that only authorized users have the ability to change, move, or delete certain types of data files. When data have integrity, they are considered accurate and can be relied upon for decision making.[8]

Availability means ensuring that data are available when needed and that IT systems are operating reliably. Stakeholders can ensure data availability in a number of ways, such as designing IT systems that are "redundant" (e.g., installed in such a way that a failure of one component will not cause an entire system to fail) and resistant to attacks, as well as ensuring that users back up data regularly.

Common intentional and malicious information security threats to IT systems and data include malware, spyware, keystroke loggers; backdoor access to IT systems; phishing and targeted scams designed to steal user credentials; intentional misuse by someone with legitimate access; and denial of service attacks intended to make data unavailable. In addition to intentional and malicious threats, those who manage IT systems and the data contained in them must protect them from unexpected or accidental events: natural disasters, power outages, and lost or misplaced IT resources (e.g., a lost thumb drive containing sensitive data). They also must protect data and related systems from the unintentional actions of legitimate users, such as the accidental deletion of important data, accidentally posting sensitive data to a public-facing resource (e.g., a web page), or sending it to the wrong person (e.g., via email).

What Is Privacy?

The rise in the ease of collecting and storing data in IT systems has also led to increased concerns about the privacy of those data. *Privacy* is a simple term used to explain concepts that apply both to individuals and to society at large. For individuals, privacy means the right of an individual to control his or her own data and to specify how those data are collected, used, and shared. In the United States, there is also a societal notion of privacy that limits the government's power to interfere in the autonomy of its citizens.[9]

Both of these concepts are important to the discussion of the national postsecondary education data ecosystem. When research practitioners initially collect student-level data, they should collect that data lawfully and in ways that seek to affirm the individual student's privacy rights. This means that collecting entities should seek permission whenever possible before collecting student data and should only collect the minimum amount of data necessary to achieve specific research goals. When these data are shared among entities in the ecosystem, particularly between institutions and state and federal entities, then they should be shared in ways that limit the government's power to use student-level data for purposes other than those first specified when the data was collected or those specifically permitted by law or regulation.

Since privacy has its roots in legal concepts, both federal and state laws may affect the entities collecting and sharing data in the national postsecondary education data ecosystem. Higher education institutions in particular are subject to a complex data protection regulatory landscape. A keen understanding of the permissions and prohibitions of some federal laws, such as the Family Educational Rights and Privacy Act of 1974 (FERPA), is crucial to ensuring student data

SIDEBAR 1: FEDERAL DATA PROTECTION LAWS

The following federal laws apply to how *institutions and non-governmental agencies* collect and use data:

The Family Educational Rights and Privacy Act of 1974 (FERPA)[10] is designed to protect students and their families by ensuring the privacy of student educational records. *Educational records* are agency or institution-maintained records containing personally identifiable student and educational data. FERPA applies to primary and secondary schools, colleges and universities, vocational colleges, and state and local educational agencies that receive funding under any program administered by the U.S. Department of Education. It does not apply to the federal government.[11] FERPA contains provisions specifying how access, amendment, and disclosure of education records must be handled. At the time of this writing, FERPA does not contain specific information security standards that institutions and agencies must use to protect student educational records.[12]

The Heath Insurance Portability and Accountability Act of 1996 (HIPAA)[13] requires covered entities (typically medical and health insurance providers and their associates) to protect the security and privacy of health records. This law is often implicated in conversations about student data when institutions have a campus medical center and student medical records are integrated with student educational records (which are protected under FERPA).

The Gramm Leach Bliley Act (GLBA)[14] applies to financial institutions and contains privacy and information security provisions that are designed to protect consumer financial data. This law also applies to how institutions collect, store, and use student financial records (e.g., records regarding tuition payments and/or financial aid) containing personally identifiable information.

The Fair and Accurate Credit Transaction Act of 2003 (FACTA or "Red Flags Rule")[15] requires entities engaged in certain kinds of consumer financial transactions (predominantly credit transactions) to be aware of the warning signs of identity theft and to take steps to respond to suspected incidents of identity theft. Like GLBA, this law applies to how institutions collect, store, and use student financial records.

The following laws apply to how *the federal government* collects and uses data:

The Privacy Act of 1974[16] is designed to protect the privacy of records created and used by the federal government. The law states the rules that a federal agency must follow to collect, use, transfer, and disclose an individual's personally identifiable information. The act also requires agencies to collect and store only the minimum information that they need to conduct their business. In addition, the law requires agencies to give the public notice about any records that it keeps that can be retrieved using a personal identifier (e.g., name or a Social Security Number [SSN]). This notice is called a system of records notice.[17] This notice describes the data being collected and how they will be used. Under the Privacy Act, the federal government cannot disclose any of the data it collects about an individual unless the underlying individual gives consent or the disclosure is made pursuant to one of 12 broad statutory exemptions. These exemptions include data disclosures made for statistical purposes to certain other federal agencies, for law enforcement purposes, and for routine uses within a federal agency. Any postsecondary education data infrastructure option that includes federal agencies collecting records with personally identifiable information will need to adhere to the requirements of this law.

E-Government Act of 2002[18] requires federal agencies to review and assess the privacy risks to their IT systems and publicly post privacy notices about their data collection practices. This law complements the Privacy Act of 1974 and was intended to promote access to electronic government resources. Under this law, an agency that collects personally identifiable information must conduct a privacy impact assessment before it collects that information.[19] The privacy impact assessment must specify the data the agency will collect, how it is collecting those data, how it will use and/or share the data, whether individuals have the opportunity to consent to specific uses of the data (e.g., any use not otherwise permitted by law), how the agency will secure the data, and whether the data collected will reside in a system of records as defined by the Privacy Act. Any postsecondary education data infrastructure option that includes a federal agency collecting data in its IT systems will need to adhere to the requirements of this law.

The Federal Information Security Management Act of 2002 (FISMA)[20] is designed to protect the security of federal information technology systems and the data contained within those systems. This law and its provisions apply to federal agencies and to contractors and affiliates of those agencies (such as educational institutions that receive a grant from a government entity). FISMA requires federal agencies to implement risk-based information security programs that conform to certain national standards. It also requires those programs to be independently reviewed each year. Any postsecondary education data infrastructure option that includes data collected and stored in a federal IT system will need to adhere to the requirements of this law.[21]

privacy. Other laws involve either the protection of specific data elements or data that are collected by federal agencies and stored in federal information systems. (See **Sidebar 1**.)

In addition to the federal laws that apply to portions of the postsecondary education data infrastructure, state laws may also affect the data collection ecosystem. In 2015, 46 states introduced over 180 bills addressing various aspects of student privacy; 15 states passed 28 new laws.[22] These laws address a multitude of issues affecting student data privacy, such as the role of educational technology service providers; whether parents can opt-out of data collection; the transfer of student data outside the state in which it was collected; data breach notification provisions; and funding for state longitudinal data systems.[23] While many states limit their laws to K-12 student data, some states have enacted laws intended

to apply to postsecondary student data as well. As a result, individual state laws will need to be reviewed for applicability in state-based data collection efforts.

Information security and privacy concepts are closely related and not always mutually exclusive. For example, it is entirely possible to envision a situation where data are secure but not private. For instance, data could be stored in an IT system in a secure way, but if those data were collected without an individual's consent or without proper legal authority, then the privacy of those data as it relates to an associated individual may be compromised. Conversely, data could be collected with an individual's consent or pursuant to a law that allows data collection, but stored in an IT system that lacks the security sufficient to protect the data from criminals seeking to steal those data. In this instance, the security of the collected data is compromised. Information security and privacy concepts must be considered together in protecting the data collected, stored, transmitted, and analyzed within the national postsecondary education data ecosystem. Privacy concepts must be followed to ensure individual and societal notions of privacy are respected to the fullest extent practicable, and information security concepts must be used to protect the confidentiality, integrity, and availability of any collected data.

Technology in the Postsecondary Education Data Infrastructure

While the security and privacy concerns and best practices described in this paper apply generally to all IT infrastructure environments, this paper series, *Envisioning the National Postsecondary Data Infrastructure in the 21st Century*, focuses on two basic types of IT system architectures for the postsecondary education data ecosystem. The first type is a single destination system, such as a student unit record data system (SURDS) operated by a governmental or private entity. As its name implies, a single destination system serves as the sole location for authoritative student information. This type of system would be operated most likely by a single entity that could impose security and privacy protections on the IT system that houses the data set. This system may receive data from a number of inputs, but the system itself is considered the single source of authoritative information.

The second type is a multiple destination system in which a number of entities share student information among themselves and no single IT system is an authoritative source of data. An example of this architecture would be several state longitudinal data systems (SLDS) networked together to respond to queries about students. In order to provide meaningful information, data must be contributed to this system in a way that allows for matching, by some common key or identifier, throughout the entire infrastructure. Given that multiple entities will contribute, use, and analyze data within the infrastructure, data security and privacy protections will need to be agreed to and followed by all participating entities throughout all contributing IT systems. **Table 1** offers a simplified view of infrastructure system architectures discussed in this paper series.

TABLE 1: SIMPLIFIED VIEW OF INFRASTRUCTURE SYSTEM ARCHITECTURES DISCUSSED IN THIS PAPER SERIES

Single destination systems	Multiple destination systems
Student unit record data system (also called student level data system)	Multiple, linked state data systems
National Student Clearinghouse	Multiple, linked federal data systems
Integrated Postsecondary Education Data System (IPEDS)	

Each of these general infrastructure architectures will require its own set of tailored security and privacy controls. For instance, creating a single destination system like a SURDS solution will require a different set of controls as opposed to bolstering systems already in place such as linking multiple SLDS systems. Although both options will require mechanisms to match records across different sources of data input, a SURDS solution could become a single, large collection of identifiable data. The information security concept of confidentiality will be highly important to this type of data collection, where tightly controlled, role-defined access will be necessary to secure a single source of stored data. A SLDS solution, on the other hand, may require more inquiry into the information security concept of integrity, for which matching data elements, resolving conflicts, and addressing quality concerns across multiple disparate systems will be necessary.

Information Security and Privacy Concerns for Postsecondary Data Collection

No matter the infrastructure architecture employed, any national postsecondary education data system will be a large collection of data designed to provide useful and reliable information about postsecondary student success and outcomes. "Big data" is a term used to define large, complex electronic data sets, usually gathered from multiple sources, as well as the transactional data (or metadata) of these data sets, which in turn must be "integrated, correlated, or otherwise analyzed together."[24] Due to their size and complexity, these data sets require concerted inquiry into how information security and privacy will be maintained across and within the data set and the IT systems that contribute to the big data collection. Due to pressing national-level questions regarding student equity and outcomes as well as the amount of data required to answer those questions, any national postsecondary education data ecosystem solution will have heightened security and privacy concerns that are similar to other big data collections.

Heightened security and privacy concerns about big data include:

- Volume (the amount of data that is collected)
- Sensitivity (the sensitivity of the data elements collected and potential sensitivity variations between different systems)
- Access (the persons or entities with access to big data collections for the purposes of querying the larger collection)

With careful planning, stakeholders can implement comprehensive information security and privacy practices into the national postsecondary education data infrastructure in a way that addresses these concerns.

Volume

The volume of personal data being collected, stored, and used by multiple parties is a common concern with big data. Volume involves two aspects. The first is the number of records included in a big data collection, whether contributed from institutions, agencies, or other organizations. This question asked in this context is "How much information do you have?" The second aspect is the number of data elements collected per individual (e.g., "How much do you know about me?"). The range of complex analytics that can be performed against a large data set, and the trends about individuals or groups that researchers can discern through such analytics, may be quite extensive. Such large collections of data, and the insights that they can reveal, are often viewed with suspicion, particularly if personally identifiable information is included among the collected data. The very act of matching data from different sources may also lead to the creation of a new data set that contains sufficient data elements to identify underlying individuals; in such instances additional privacy protections are required to protect individuals.[25]

In the national postsecondary education data ecosystem, both volume issues are implicated. Data may be available from multiple entities (e.g., institutions, state and federal agencies) throughout the ecosystem. In addition, different entities (and their underlying IT systems) may hold different individual data elements. When these data elements are combined for analysis, the results may yield a more detailed picture about an individual than any of the entities could have drawn on their own.

Sensitivity

The sensitivity of data in big data collections is also a concern. There are two different, but very closely related, sensitivity issues. The first is acknowledging that big data sets may include different categories of data in any number of combinations. Common types of data categories are identifiable, de-identified, anonymous, and aggregate data. (See **Sidebar 2**.) Different categories of data may need markedly different security and privacy protections because of the sensitivity of that data. For instance, identifiable data categories are typically considered to be more sensitive than anonymous data elements.

The second sensitivity issue is closely related to the first but includes a nuance with respect to the use of identifiable data in a big data collection. Not all types of identifiable data are equal in terms of their sensitivity. Understanding the sensitivity of different elements in the data set is crucial. Some data elements, like name and Social Security Number (SSN), are among the most sensitive types of data because they alone can identify a unique individual or because they are considered to be highly sensitive by societal standards.[26] Often state or federal laws such as FERPA or HIPAA protect these most sensitive data elements.

Some data elements may be considered sensitive because, when combined, they are more likely to identify a unique individual. Still other elements are considered personally identifiable data, but they are less likely to be used to identify a unique individual. **Table 2** highlights some different types of personally identifiable data elements and their sensitivity.

When the different data categories and personally identifiable data elements are combined in a big data collection, and queries are returned with differing levels of data element sensitivity, the results of those queries must also be secured in a manner that protects the most sensitive data returned in that query. This can be a challenge in both single destination and multiple destination data systems.

In the national postsecondary education data ecosystem, practitioners will need to address sensitivity challenges in two ways. First, the entities participating in the ecosystem must implement the appropriate security and privacy protections in IT systems under their control. This means making sure that the data collected in those systems is protected at a level that corresponds to the most sensitive data element residing in those systems. Different systems, even within a single entity, may require different levels of protection. In addition to individual IT system level protections, all entities within the national postsecondary education data ecosystem will need to work together to ensure that data shared amongst entities and within the ecosystem is properly protected at the endpoint, where it is ultimately reported to an end user, according to the most sensitive data element included in any analysis.

TABLE 2: PERSONALLY IDENTIFIABLE DATA ELEMENTS

	Most Sensitive	More Sensitive	Less Sensitive
Summary	Data elements that directly identify a unique individual	Data elements that, when combined, may identify a unique individual	Data elements that are unlikely to identify a unique individual
Data elements	Name Social Security Number Driver's license number Passport number Taxpayer identification number Biometric information (e.g. fingerprints, retina scans, voice signature, facial geometry) Photographic images of the face or distinguishing characteristics	Physical address Email address Telephone number Date of birth (becomes "most sensitive" when paired with a name or parts of a name) Financial account numbers and financial transaction information Health information and health insurance information Employment information Educational information (e.g., schools attended, enrollment dates, graduation date, academic preparation level, academic interactions between institution and student) Demographic characteristics such as gender, race, ethnicity, religion	Zip code Telephone area code

Access

The widespread availability of personal data being collected and stored, and the near global availability of data through the Internet and via personal mobile devices, poses unique access concerns for big data collections. These collections are almost always designed to be accessed by multiple entities, from various locations, and for different purposes. Access concerns generally fall into two categories: 1) protecting data from external actors without legitimate access to the data; and 2) protecting data from internal actors such as individuals with legitimate access to systems who intentionally exceed the scope of their pre-approved authority, who access data from unapproved devices, or who make mistakes and accidentally disclose data.[27] In the national postsecondary education data infrastructure, in which multiple IT systems could be linked, data sets from multiple owners are combined into shared systems, and many individuals may need access to the data, information security controls are necessary to not only secure the data from external intrusion but also to implement access control policies for legitimate individuals.

Information Security and Privacy Protections in the Postsecondary Education Data Infrastructure

There is no one-size-fits-all formula for designing information security and privacy protections that ensures the security and privacy of all data moving within a postsecondary education data system and the underlying IT systems that form the entire ecosystem. Each of the options explored in *Envisioning the National Postsecondary Data Infrastructure in the 21st Century* will offer its own set of unique information security and privacy technology challenges that must be individually addressed based on the underlying technologies and processes used to provide the solution. Thus, a holistic approach that favors best practices, overall risk reduction, data safeguards, and transparency, accountability, and trust throughout the entire ecosystem is essential.

Information Security Protections

A number of resources for information security standards and best practices exist. (See **Sidebar 3**.) Almost all of these standards are based on the concept that good information security practices attempt to *reduce risk* and *safeguard data*. In this case, *risk* is the likelihood that a threat will exploit a vulnerability to cause harm. An example of a risk would be a malicious hacker (the threat) guessing a user's weak IT system password (a vulnerability) to steal data from a database that is later used to cause harm to an individual (identity theft). The likelihood of any risk being realized and the impact or harm of that risk being realized varies from context to context. Not all risks or vulnerabilities require the same level of attention, and most organizations do not have the resources necessary to attempt to eliminate all information security risk. Finding a way to discern between different types of risks and assess their relative criticality, called risk assessment, must be an essential component of the national postsecondary education data infrastructure.

Most risk management methodologies[28] include four basic risk assessment steps:

1. Inventorying the IT assets and data included in the scope of the assessment
2. Identifying the threats and vulnerabilities to those assets and data (collectively called risks)

SIDEBAR 2: KNOW THE LINGO

When thinking about big data, it is important to understand the different types of data that can be included in a data set:

Personally identifiable data identifies a specific individual and is also known as personally identifiable information or PII. Personally identifiable data can include a single piece of information used alone, such as a person's name or SSN, or data elements that, when combined, can identify a particular individual. Common personally identifiable data elements include name, SSN, physical address, email address, zip code, race, age, gender, GPS location, telephone number, college or university identification number, and account numbers. (See **Table 2**.)

Aggregate data includes individual data elements that have been combined to statistically analyze trends. Aggregate data protects privacy by categorizing individuals with similar characteristics rather than isolating a single individual. To effectively aggregate data so that they are difficult to re-identify individuals, the data set should include a large population of individuals who are categorized in such a way that broad sets of individuals are classified together and the data set does not include extra information that would be unique to a single individual in the data set. Entities may place limits on how data can be aggregated to protect privacy in a defined population. For instance, an entity might not report aggregate data for a population that has less than a certain number of participants because it still could be easy to infer information about individuals in that population even if data were aggregated.

De-identified data have had all personally identifying information removed in order to protect privacy. De-identified data are not necessarily the same as anonymized data. Depending on how the de-identification is accomplished, personally identifying information may be able to be re-associated with the data set at a later time.

Anonymized data can no longer be associated with an individual in any way. The act of permanently and completely removing personal identifiers from data is called anonymization. Data are anonymized when stripped of personally identifying elements and those elements can never again be re-associated with the original data or the underlying individual. Often de-identified and anonymized data are considered the same, but understanding the nuance between the two types of data is useful when establishing technical requirements for security and privacy in large data sets.

3. Categorizing the likelihood of and the potential loss from certain risks occurring
4. Documenting where controls are needed to address the identified risks

The major outcome of a risk assessment is to identify the risks to IT assets and data according to a matrix based on likelihood and impact (e.g., low, medium, and high) and to develop a plan for addressing those risks in a way that makes sense for the underlying organization. Depending on the organization's risk tolerance, it may choose to address a) the risks that are most likely to occur, b) the risks most likely to cause the greatest loss if they occur, c) the most easily addressed risks from a resources perspective, or d) the risks that involve some combination of all the above.

After practitioners identify and assess a particular risk, they can apply information security controls to address it. The ultimate goal of assessing risk and applying information security controls is to protect an organization's IT resources and the data contained within those resources. A number of controls from the following general information security areas must be used to properly address risk:

- **Asset management** focuses on how IT systems and the data within those systems are managed throughout their lifecycle from creation or acquisition to destruction.
- **Identification, authentication, and access control** relate to how authorized users are identified, authenticated (prove their identity), and given access to IT systems and/or data within those systems.
- **Operational security** pertains to how IT systems and the data contained within them are operated, protected from threats, and tested for vulnerabilities. Malware protection, system logging and monitoring, data backups, and vulnerability management are all included in this general category.
- **Communications security** pertains to how IT systems and the data within those systems are protected when data move between networks or across IT systems, both within a single organization or among multiple organizations.
- **Physical and environmental security** relate to how IT systems and the data contained within them are protected from physical loss, mechanical failure, and environmental damage. This includes protecting IT systems from risk of theft or loss; natural disasters such as fires, floods, and tornados; intentional vandalism; and power loss or other mechanical failure.
- **Incident response, business continuity, and disaster recovery** pertain to how organizations respond to incidents involving IT systems and the data contained within those systems, and how organizations recover from those incidents. An organization must implement response and recovery protocols for a number of different types of incidents (e.g., a malicious attack, a natural disaster, or intermittent loss of Internet connectivity).
- **Training and awareness** relate to how organizations train their employees and other IT users and spread awareness about how the organization promotes good information security practices. Training and awareness are important because employees and other trusted individuals, even with the best of intentions, can unknowingly compromise the security of IT systems and the data contained within those systems.

Privacy Protections

As a subject matter area, privacy has grown in importance in higher education over the past 10 years.[29] Laws including FERPA and the growth of the educational technology market have made privacy concepts even more important within the education ecosystem. Much like information security concepts, there is no single comprehensive set of controls to ensure adequate data privacy in all situations, and each possible data solution will present unique privacy challenges requiring a specialized privacy response.

SIDEBAR 3: INFORMATION SECURITY CONTROLS

A number of resources specify information security controls that organizations can implement to protect their IT systems and the data within those systems. Two of the most common guides are the National Institute for Standards and Technology (NIST) Special Publication 800-53 (Rev. 4), *Security and Privacy Controls for Federal Information Systems and Organizations* (2015),[30] as well as the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) joint publication ISO/IEC 27002:2013, *Information Technology-Security Techniques-Code of Practice for Information Security Controls* (2013).[31]

NIST, an agency located within the U.S. Department of Commerce, creates information security guidance for federal agencies. Federal agencies are required to follow NIST guidelines as part of their FISMA compliance activities. Non-federal organizations, at their discretion, may use NIST guidelines as best practices to improve their own information security programs.

NIST Special Publication 800-53 states the minimum security and privacy controls that federal agencies should follow to secure federal IT systems. First published in 2005 and revised four times, NIST Special Publication 800-53 is essentially a catalog of security and privacy controls that can be implemented within federal agencies' IT systems to protect those systems from a number of different risks such as hostile attacks or human error. There are 18 families of security controls listed in the publication and within each family is guidance on how to select appropriate safeguards to meet an organization's information security goals.

The International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) are two independent non-governmental organizations that collaborate to strengthen standards systems. The standards issued by these organizations are consensus-based and are used as references in international trade. First published in 2005 and most recently revised in 2013, *Information Technology-Security Techniques-Code of Practice for Information Security Controls* (ISO/IEC 27002:2013) is a practical guide for implementing security controls and includes 14 major categories of controls. Although its use is not mandated under U.S. law, ISO/IEC 27002:2013 is often cited as one of the main set of discretionary controls for securing information technology systems.

Table 3 compares the NIST and ISO/IEC families of security controls. There is significant overlap between the two lists. In fact, NIST Special Publication 800-53 specifically includes an appendix that maps NIST controls to ISO/IEC 27002:2013 controls to illustrate their overlap.

TABLE 3: CONTROL FAMILIES FOR NIST 800-53 AND ISO/IEC 27002:2013

NIST 800-53 Security Control Families	ISO/IEC 27002:2013 Security Control Categories
Access Control	Asset Management
Audit and Accountability	Access Control
Awareness and Training	Business Continuity Management
Configuration Management	Communications Security
Contingency Planning	Compliance
Identification and Authentication	Cryptography
Incident Response	Human Resources Security
Maintenance	Incident Management
Media Protection	Information Security Policies
Personnel Security	Operations Security
Physical and Environmental Protection	Organization of Information Security
Planning	Physical and Environmental Security
Program Management	Supplier Relationships
Risk Assessment	System Acquisition, Development, and Maintenance
Security Assessment and Authorization	
System and Services Acquisition	
System and Communications Protection	
System and Information Integrity	

Nonetheless, establishing a shared a set of privacy principles will be instrumental in helping protect student privacy across all options for improving the national postsecondary education data infrastructure. Adherence to these principles will promote transparency, accountability, and trust within the national ecosystem. The *Fair Information Practice Principles* (FIPPs) are part of the federal Privacy Act of 1974 and have been influential in shaping U.S. privacy law. They were designed to address privacy concerns rising from larger digital collections of personal data and thus provide a good model for the privacy principles to be adopted within the national postsecondary education data infrastructure. The FIPPs consist of eight privacy principles:

1. **Purpose Specification:** Organizations should tell individuals why data are being collected, and for which uses, before the data are actually collected.
2. **Collection Limitation:** Organizations should collect only the data that they need (called data minimization) and must obtain those data by lawful means or with notice and consent from the associated individual.
3. **Data Quality:** Organizations should only collect accurate data and should have a process (called redress) that an individual can follow if for some reason the data about that individual are not correct.
4. **Use Limitation:** Organizations should only use data for the purposes specified when the data are originally collected or otherwise permitted by law.
5. **Security Safeguards:** Organizations should protect collected data from unauthorized access (e.g., confidentiality), destruction (e.g., availability), and modification (e.g., integrity).
6. **Openness:** Organizations should be transparent and provide individuals with information about their data collection activities.
7. **Individual Participation:** Individuals should be able to find out if data about them have been collected by an organization, and they should have access to any such data.
8. **Accountability:** Organizations that collect data must be held accountable for the aforementioned the privacy principles.

Privacy principles cannot be addressed in a vacuum. The entire postsecondary education data community must agree on the principles that will be applied across the ecosystem. To do this, stakeholders should consider a collaborative data governance process as the ecosystem evolves. A data governance program would put in place the policies and processes needed to manage the data collected, used, and shared within the national postsecondary education data infrastructure. Such a program would provide guidance on the data that are available, the sensitivity of those data, who is responsible for them, where they are stored, who has access, and the risks and regulations associated with those data.[32]

Recommendations for Policymakers

Providing reliable data to students, parents, administrators, faculty, policymakers, and other leaders interested in student outcomes and ensuring the security and privacy of those data are not mutually exclusive tasks. The following four recommendations work together to form a holistic framework for ensuring effective information security and privacy protections within the national postsecondary education data infrastructure ecosystem:

1. **Adopt a risk-based approach to understanding information security and privacy threats and vulnerabilities.** Regardless of the national postsecondary education data infrastructure solution or its architecture, stakeholders must understand the information security and privacy risks that could affect any system's ability to provide stakeholders with the information needed to improve student outcomes. After practitioners have assessed the risks, they can apply information security and privacy controls to address that risk and to secure the IT systems, and the data within those systems, that constitute the national postsecondary education data infrastructure.
2. **Establish and adhere to a baseline set of information security protections.** These protections are necessary to safeguard the data collected, processed, stored, and transmitted within the national postsecondary education data infrastructure. If a set of standards is not otherwise required by state or federal law (for example, the use of NIST Special Publication 800-53 to implement controls to protect federal IT systems), then at a minimum the controls implemented must be based upon the risks inherent in the different systems within the ecosystem (See **Sidebar 4**).
3. **Establish and adhere to a baseline set of privacy standards.** Protecting student privacy within the national postsecondary education data infrastructure requires adopting a guiding set of privacy principles. Adopting these principles before a national effort is undertaken would provide the best privacy solution for students. At a minimum these principles should require that a) individuals receive notice and provide consent before data are collected; b) that institutions and other organizations only collect the minimum data needed to answer the critical questions to feed the key student outcomes measures; and c) that institutions and other organizations only use collected data for the purposes for which they were originally collected or for purposes that are otherwise permitted by law.
4. **Implement a collaborative governance structure.** A governance structure that ensures that data collected within the national postsecondary infrastructure feeds necessary measures and metrics and answers stakeholders' questions is essential. This governance structure can also be used to review the data available within the ecosystem

and ensure that data are protected. In addition to defining data ownership and stewardship practices and advising on information security and privacy best practices and baseline requirements, a governance entity could consider how best to train users on these systems and communicate the benefits of concerted data sharing and analytics.

SIDEBAR 4: A DE FACTO STANDARD?

In June 2015, the National Institute of Standards and Technology (NIST) published Special Publication 800-171, *Protecting Controlled Unclassified Information in Nonfederal Information Systems and Organizations*.[33] The purpose of this publication is to provide guidance to federal agencies to ensure that certain types of federal information are protected when processed, stored, and used in *nonfederal* information systems. NIST Special Publication 800-171 applies to data that the federal government designates as Controlled Unclassified Information (CUI) when they are shared by the federal government *with* a nonfederal entity and when *no other* federal law or regulation (e.g., FISMA) addresses how to protect the underlying data.

In the higher education context, the federal government often shares CUI-designated data with institutions for research purposes or in carrying out the work of federal agencies. It should be noted that *student records or personally identifiable information* are considered CUI.[34] As such, the controls specified in NIST Special Publication 800-171 (which are based on NIST Special Publication 800-53, mentioned previously in this paper) will need to be addressed in any higher education IT system that stores or processes any CUI received from the federal government pursuant to a contract with a federal agency.

Institutions are still trying to understand the impact of NIST Special Publication 800-171 on their IT systems and the data they receive from the federal government. As the national postsecondary education data infrastructure evolves, it is entirely possible that the provisions of NIST Special Publication 800-171 will create a *de facto* information security standard for this ecosystem.

Conclusion

Students, institutions, and policymakers need better information about postsecondary education. Better data, retrieved through both existing initiatives and as well as through options currently under consideration, are required to provide the meaningful information about student outcomes that these stakeholders need most. As stakeholders consider the best ways to meet these data needs, it is imperative that their conversations include how to best protect student privacy and ensure the security of data used throughout the national postsecondary data system. With thoughtful planning, comprehensive information security and privacy practices can be implemented within the national postsecondary education data infrastructure in such a way that reduces risk, safeguards data, and ensures transparency, accountability, and trust throughout the entire ecosystem.

Endnotes

1 Voight, M., Long, A., Huelsman, M., and Engle, J. (2014). *Mapping the postsecondary data domain: Problems and possibilities*. Washington, DC: Institute for Higher Education Policy. Retrieved from http://www.ihep.org/research/publications/mapping-postsecondary-data-domain-problems-and-possibilities; Engle, J.. (2016) *Answering the call: Institutions and states lead the way toward better measures of postsecondary performance*. Seattle: Bill and Melinda Gates Foundation. Retrieved from http://postsecondary.gatesfoundation.org/wp-content/uploads/2016/02/AnsweringtheCall.pdf

2 Voight et al., *Mapping the postsecondary data domain*.

3 Engle, *Answering the call*.

4 Rorison, J., and Voight, M. (2015). *Weighing the options for improving the national postsecondary education data infrastructure*. Washington, DC: Institute for Higher Education Policy. Retrieved from http://www.ihep.org/research/publications/weighing-options-improving-national-postsecondary-data-infrastructure

5 In this paper, the term ecosystem is used to refer to the various national postsecondary education data infrastructure reform options and the underlying IT systems, including federal, state, local, and/or institutional, that exist to collect, store, transmit, and analyze data within and across each option.

6 Gorman, A., and Sewell, A. (2015, July 12). Six people fired from Cedars-Sinai over patient privacy breaches. *Los Angeles Times*, July 12, 2013. Retrieved from http://articles.latimes.com/2013/jul/12/local/la-me-hospital-security-breach-20130713

7 Confidentiality is also a concept used in many other contexts, such as the practice of not disclosing certain records because they contain sensitive, harmful, or embarrassing information. In this paper, however, the term confidentiality specifically refers to a technical information security concept unless otherwise indicated.

8 Information security concepts of integrity are often implicated in larger discussions about data quality.

9 Societal notions of privacy stem from the U.S. Constitution, particularly the Fourth Amendment, which protects individuals against unreasonable government searches and seizures. This notion is also present in federal laws that seek to place limitations on how government agencies can use data.

10 Family Educational Rights and Privacy Act of 1974 (FERPA), U. S. Code, vol. 20, sec. 1232g (2012).

11 Note, however, that FERPA does contain provisions on what types of education records can be shared with the federal government and how that data can be subsequently used.

12 A bill that would overhaul FERPA privacy protections was introduced in July 2015. The Student Privacy Protection Act (H.R. 3157) is intended to modernize privacy protections, improve communication, and "hold schools, states and independent entities accountable for their use of student information." Despite its focus on student privacy, FERPA as currently written does not include any data security provisions. At the time this writing, the proposed act includes language that would mandate that institutions implement data protection policies (that are assumed to include security practices) and notify students in the event of a data breach. A fact sheet on the legislation can be found here: http://edworkforce.house.gov/uploadedfiles/fact_sheet_-_student_privacy_protection_act.pdf.

13 Health Insurance Portability and Accountability Act of 1996 (HIPAA), U.S. Code, vol. 42, sec. 1320d (2012).

14 Gramm Leach Bliley Act (1999) (GLBA), Title V of the Financial Services Modernization Act of 1999, U.S. Code, vol. 15, sec. 6801, *et seq.* (2012).

15 Fair and Accurate Credit Transaction Act of 2003 (FACTA), Pub L. 108-159, 117 Stat. 1952.

16 Privacy Act of 1974, U.S. Code, vol. 5, sec 552a (2012).

17 The U.S. Department of Education's Privacy Act System of Record Notice issuances for records that it collects are available at: http://www2.ed.gov/notices/ed-pia.html.

18 E-Government Act of 2002, Public Law 107-347, codified in scattered sections throughout U.S. Code, vol. 44 (various sections), (2012).

19 The U.S. Department of Education's Privacy Impact Assessments for its departmental IT systems are available at: http://www2.ed.gov/notices/pia/index.html.

20 Federal Information Security Management Act of 2002 (FISMA), Title III of the E-Government Act of 2002, U.S. Code, vol. 44, sec. 3541 et seq. (2012).

21 The U.S. Department of Education's Federal Student Aid Office FISMA statement can be found at: https://studentaid.ed.gov/sa/privacy#security.

22 Data Quality Campaign. (2015). Student data privacy legislation: What happened in 2015, and what is next? Retrieved from http://dataqualitycampaign.org/wp-content/uploads/2015/09/Student-Data-Privacy-Legislation-2015.pdf

23 Data Quality Campaign, Student data privacy legislation.

24 Almes, G.T., Hillegas, C.W., Lance, T., Lynch, C.A., Monaco, G.E., Mundrane, M.R., and Zottola, R.J. (2014). *Big data: Laying the groundwork*. Louisville, CO: EDUCAUSE Center for Analysis and Research. Retrieved from http://www.educause.edu/library/resources/big-data-in-the-campus-landscape

25 Barnett, W., Corn, M., Hillegas, C., and Wada, K. (2015). *Big data in the campus landscape: Security and privacy*. ECAR working group paper. Louisville, CO: EDUCAUSE Center for Analysis and Research. Retrieved from http://www.educause.edu/library/resources/big-data-in-the-campus-landscape

26 EDUCAUSE. (2014). *Managing data risk in student success systems: EDUCAUSE IPAS summit report*. Louisville, CO: EDUCAUSE. Retrieved from https://net.educause.edu/ir/library/pdf/PUB9015.pdf

27 At higher education institutions, from 2005-2013, the most common cause of data breaches was electronic entry by an outside party (36 percent), followed by unintended disclosures committed from within an institution (30 percent). See Grama, J. L. (2014). *Just in time research: Data breaches in higher education*. Louisville, CO: EDUCAUSE Center for Analysis and Research. Retrieved from http://www.educause.edu/library/resources/just-time-research-data-breaches-higher-education

28 There are a number of risk management methodologies that can be used to assess risks to information technology systems and data. NIST has developed a Risk Management Framework for federal agencies to follow as part of their FISMA obligations. See http://csrc.nist.gov/groups/SMA/fisma/framework.html.

29 Vogel, V.M. (2015). The chief privacy officer in higher education. *EDUCAUSE Review*. Retrieved from http://er.educause.edu/articles/2015/5/the-chief-privacy-officer-in-higher-education

30 Joint Taskforce Transformation Initiative. (2015). *Security and privacy controls for federal information systems and organizations*, Special Publication 800-53, Rev. 4. Washington, DC: National Institute of Standards and Technology. Retrieved from http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r4.pdf

31 International Organization for Standardization and the International Electrotechnical Commission. (2013). *Information technology-Security techniques-Code of practice for information security controls*, ISO/IEC 27002:2013. Geneva: International Organization for Standardization. Retrieved from http://www.iso.org/iso/home/store

32 Blair, D., Briner, K., Dani, V., Fary, M., Fishbain, J., Hart, M.D., Hopkins, B.W., Kelly, M.C., Matuch, K., Zaborowski, E. (2015). *The compelling case for data governance*. Louisville, CO: EDUCAUSE Center for Analysis and Research. Retrieved from http://www.educause.edu/library/resources/compelling-case-data-governance.

33 Ross, R., Viscuso, P., Guissanie, G., Dempsey, K., Riddle, M. (2015). *Protecting controlled unclassified information in nonfederal information systems and organizations*, Special Publication 800-171. Washington, DC: National Institute of Standards and Technology. Retrieved from http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-171.pdf

34 See the Controlled Unclassified Information (CUI) registry for student records at: http://www.archives.gov/cui/registry/category-detail/privacy-student-records.html.

Envisioning the National Postsecondary Data Infrastructure in the 21st Century is a project of the Institute for Higher Education Policy and is supported by the Bill & Melinda Gates Foundation.