Implementing a Student-Level Data Network (Part IV):

EXPLORING DESIGN OPTIONS

James Isaac and Erin Velez, RTI International[1]
Amanda Janice Roberson, Institute for Higher Education Policy

RTI INTERNATIONAL

IHEP

BACKGROUND

The current federal postsecondary education landscape includes high-quality data that help constituents address questions related to students and their outcomes, but gaps persist that leave critical questions from the higher education community unanswered. To expand the availability and coverage of federal postsecondary education data, Congress introduced the College Transparency Act ([CTA]; U.S. Senate and U.S. House of Representatives [House]) in the 115th through 117th Congresses and the College Affordability Act ([CAA]; House) in the 116th Congress. In 2022, the House passed the CTA as an amendment to the America COMPETES Act, a larger bill focused on workforce competitiveness, though the amendment was excluded from the final bill that was signed into law.

The mechanism for comprehensive data reform proposed by the CTA and CAA is the creation of a federal student-level data network (SLDN) that would leverage data available at institutions of higher education and federal agencies with the intent of streamlining those institutions' data-reporting burden and ensuring that reported data are inclusive of all students and all outcomes. Legislation includes detail on the data elements, data coverage, agencies responsible for data collection, and advisory body but is not prescriptive on the design and implementation of an SLDN.

To help create a knowledge base around an SLDN in preparation for passage of the CTA, RTI International[2] partnered with the Institute for Higher Education Policy (IHEP) to host a series of forums on the SLDN. The first SLDN forum was held in June 2020 and sought to outline the data elements required by the legislation.[3] In September of that same year, RTI and IHEP convened a second forum regarding institutional perspectives on the data submission process to an SLDN.[4] Panelists in the first two forums raised issues related to financial aid variables in an SLDN, which precipitated a third forum to engage financial aid professionals in May 2021.[5] A fourth forum, held in May 2022, is the focus of this report and was centered around three potential models for a federal SLDN, each separately devised by pairs of experts knowledgeable in student-level postsecondary data collection efforts. Participants attending the forum provided feedback and areas for further exploration related to each design.

The CTA requires the SLDN be a student-level data collection for all postsecondary students at every institution in the country that is eligible to participate in federal financial aid programs, which covers about 20 million students annually across almost 6,000 institutions.[6] The National Center for Education Statistics (NCES), at the U.S. Department of Education (ED), would create and implement the SLDN, including the collection and reporting by institutions of approximately 40 data elements per student, with the expectation that roughly 20 additional data elements per student would be included from federal data sources (e.g., National Student Loan Data System, Internal Revenue Service, Census Bureau, Veterans Benefits Administration, and U.S. Department of Defense).

RTI and IHEP plan to continue to engage experts and gather intelligence, supporting the brainstorming around SLDN design in preparation for future passage of the CTA or a similar bill creating an SLDN.

The information summarized in this report is intended to generate discussion and is not intended to draw conclusions or recommend solutions for the SLDN. As specified in the CTA, the Commissioner of NCES is tasked with developing and maintaining the secure, privacy-protected SLDN in a manner that aligns with existing federal standards of privacy and web design. RTI and IHEP plan to continue to engage experts and gather intelligence, supporting the brainstorming around SLDN design in preparation for future passage of the CTA or a similar bill creating an SLDN.

PROCESS

Beginning in late 2021, RTI met with more than a dozen experts who had worked to develop state-level student data systems, contributed to federal education data systems, installed privacy safeguards for university systems, or managed institutional data collection efforts. The listening tour helped RTI identify and engage six professionals, in three teams of two, to develop high-level conceptual plans regarding the design and implementation of a federal SLDN. See Appendix A for brief biographies of each team member. Teams presented their models during the forum, which included 18 panelists (listed in Appendix B). Each team also provided working papers that further articulated the details of each model, and these were provided to all panelists ahead of the forum. Appendix C includes the three working papers.[7]

RTI staff served as moderators for the forum, encouraging panelists to engage with the models and discuss them individually and collectively. Moderators and presenters paid special attention to benefits and limitations of each design when realistically considering the development of an SLDN. Forum attendees also discussed steps that NCES and ED could take to facilitate a field-informed design of the SLDN.

DISCUSSION

The following is a review of central SLDN components addressed by each model and related topics discussed during the forum. The following summary differentiates panelists—those who attended the forum and posed questions—from presenters, who designed and presented the three models.

Model 1: A Model for an SLDN

Authors: Tod Massa and Michelle Appel

GOALS

- Align and integrate the SLDN with state student data collection systems or efforts
- Separate the collection of student data from the collection of institutional data
- Allow demographic data elements to be edited over time, as attributes like gender and race and ethnicity are not permanent characteristics for all students
- Allow data to give voice to populations that are unseen while protecting every individual's right to privacy

The model proposes using a process similar to one that currently exists in Virginia, wherein institutions periodically submit student data. The data submitted are matched with existing data using a registry key that is separately maintained and updated as new students enroll at any participating SLDN institution. The registry key includes both student information and institutional information (including programs offered), enabling verification of all records submitted. Student IDs are stored in the registry, which allows a unique student ID to be created once, when a student is included in a data submission for the first time. If a student is new to the institution, but not to the federal SLDN, the student's originally assigned ID is found and delivered to the institution. Social Security Numbers (SSNs) are used to assign new student IDs, but thereafter only the student ID is used for future data submissions.

Additionally, it is important that states have the ability to submit data to the federal SLDN on behalf of institutions, consistent with legislative language. The model notes that a way to manage burden is to leverage state processes to support data submission, such that verification and quality assurance could be conducted at the state level, reducing errors when data are submitted at the federal level. The state could provide technical support and help desk functions to institutions, in addition to what may already exist. A benefit of this model is that states can continue to operate on timelines required for their state-level operations, without having to wait on data provided back to them by the federal system.

The presenters noted that there is a need to accommodate institutions with small numbers of students and sparse technology resources through templates, technical support, and training, given that an SLDN must include all of the approximately 6,000 institutions across 57 states and territories that received federal financial aid. The presenters also noted that current federal privacy laws pose barriers that must be considered. For example, they posited that current privacy laws would not allow an institution to submit a list of student names and get back individual-level nondirectory information that the institution did not originally submit to answer questions about pathways and outcomes. According to the Family Educational Rights and Privacy Act (FERPA), the SLDN can only share individual-level data if they are directory information, which is considered to be in the public domain.[8] As a result, much of the reporting from the SLDN may need to be shared in the aggregate, not at the student level.

The presenters stressed that trust is essential to the success of such a system. Data governance is critical to creating structure and buy-in for the system, as are the guardrails stipulated in the CTA legislation. Another crucial feature is information justice, wherein voice is given to populations that are usually unheard (e.g., American Indian students and Native Hawaiian students), while ensuring the protection of privacy for those groups and individuals. The presenters emphasized, and panelists agreed, that a system is more likely to succeed if all constituents, including the institutions required to submit data, find benefit in the system.

Model 2: Building a Scalable Student Unit-Level Data Model for Reporting

Authors: Richard Reeves and Valeria Garcia

GOALS

- Enhance the public's ability to conduct research on student pathways and outcomes
- Provide additional analytic support to all institutions
- Offer a scalable solution to the data collection process

The presenters of Model 2 built their model around a multitiered value proposition wherein public data users, regulatory bodies, and institutions could all benefit from an SLDN that engages new technologies. The model is predicated upon a multitenant data collection system that leverages Infrastructure as a Service (IaaS) and Platform as a Service (PaaS) technologies to improve the efficiency of data collection and reporting platforms. In short, the model allows for storing each institution's data separately and providing institutions continuous access to their data submission on an analysis platform. This enables institutions to submit data to the SLDN as well as access and use the data by running analyses on the platform. Presenters also envision data users developing new metrics and statistics using these data, without creating additional burden for the reporting institutions.

The SLDN, as articulated in the CTA, must include the capacity to identify individual students and their pathways through postsecondary education and the workforce. Given this, the presenters propose assigning each student a unique National Student Identifier (NSID). The NSID would be separate from other federal identifiers, known only to the student, the educational institution, and NCES. This approach enables the database to include unique identifiers for each educational record without relying on information that might be subverted in unintended ways, which ultimately helps protect student privacy.

Presenters estimated that the cost of an SLDN could be $30 to $80 million annually, based on the cost and scale of the Integrated Postsecondary Education Data System (IPEDS) and National Student Loan Data System (NSLDS), as well as the costs of a multitenant technological system. Presenters noted that the cost would depend, in part, on whether and how much support the SLDN would provide to federal student aid administration efforts.

Model 3: Data Lakehouses and Privacy by Design: An Architectural Vision for a Proposed Federal Longitudinal Data Network

Authors: Brendan Aldrich and Pegah Parsi

GOALS

- Adopt a privacy-by-design mindset that incorporates privacy into systems, priorities, and processes to protect students and their families
- Leverage technical advancements, like data lakehouses, to store, edit, and query data in a more efficient and dynamic manner than currently used
- Accept "as of" file submissions whereby institutions can update or edit reported data, while changes are clearly recorded

The proposed legislation requires that the SLDN be privacy-protected because it would include rich, sensitive detail regarding individuals and their families—the type of information that could empower students and their families to make data-informed decisions. To protect this information, the presenters recommended a privacy-by-design model in which privacy is an essential component at the core of the SLDN.

The model envisions the SLDN as using a cloud-based approach to data storage and computation in the form of a *multilevel data lakehouse*. Presenters described this as a combination of the best elements and capabilities of traditional data warehouses and data lakes[9] within a single technological architecture. Benefits of this design include administration at scale, minimal data movement between upload and analyses, and cost-effective data storage with minimal duplication.

The multiple levels stipulated in the SLDN data lakehouse represent different potential data providers—institutions, consortia, and federal sources—and data users. Depending on their level, members have differential access to the data contained within the overarching SLDN lakehouse. At the institution level of the lakehouse, institutional representatives have access to their own raw and transformed data files, including, query-able tables, conformed data tables, and curated data tables. In contrast, federal agencies that merge data with the SLDN have prescribed access to data submitted by all institutions via the transformed and query-able curated tables.

As with the other models presented during the forum, the third model also aims to reduce institutional burden. The model achieves this reduction by accepting data in an institution's native format—along with a codebook—and then performing any necessary data conformation within the federal lakehouse. The presenters propose that moving data conformation functions to the federal level, in addition to increasing the ease of submission, improves transparency because decisions are made by a single entity rather than on an institution-by-institution basis. This model also advocates for an "as of" approach to data submission, in which data are treated in an additive manner and historical errors are rectified by subsequent data uploads.

Lessons Learned and Need for a Federal SLDN

Presenters and panelists who participated in this forum have extensive experience working with and examining topics related to postsecondary education data collection at the state and federal levels. Familiarity between presenters and panelists promoted an open dialogue throughout the discussion.

As one example, presenters and panelists discussed lessons learned over many years grappling with the notion of a federal SLDN. One presenter explained that this modeling exercise allowed for the exploration of the importance of data privacy and the power that data can have for groups of people. Specifically, this presenter noted that a federal SLDN could give voice to people who are otherwise overlooked and underserved by the postsecondary system, noting that data could be examined by gender identity or other demographic details. The panelist expressed awareness of risks associated with government entities holding personal student data. Presenters and panelists agreed that SLDN data collection and distribution must walk a very fine line to benefit students and those whose data are collected in a manner that does not put them at risk—data in an SLDN are meant to help students, never to harm them, and the designs discussed at this convening seek to strike this balance.

Data in an SLDN are meant to help students, never to harm them, and the designs discussed at this convening seek to strike this balance.

Regarding the comparison of a hypothetical SLDN versus current federal data collection efforts, a presenter noted that IPEDS has an extensive glossary of terms and a useful framework for data governance. Additionally, the transition away from IPEDS to an SLDN would require time, funding, and effort. One presenter was asked directly whether the transition would be worthwhile. The presenter responded that, given that higher education is a huge part of the nation's economy and a huge driver of social mobility, a federal SLDN would be worth the investment. An SLDN offers the much-desired opportunity to measure outcomes of postsecondary degree attainment and repayment of student loans on individual factors such as socioeconomic mobility, continued education, and employment. As such, if the SLDN can provide researchers the data to analyze how postsecondary education can support broader efforts to increase social and economic mobility, then it would be worth the cost and effort to transition. Other presenters and panelists voiced their agreement.

Student Identifiers

Panelists questioned the presenters on methods to identify students, merge data between federal sources, and maintain individual privacy. Discussion of these questions built upon specifics addressed in each model, such as the use of SSNs to create a student identifier number or other options, including those known to be applied by current federal and state data collection efforts.

One presenter noted that a student's name, date of birth, and zip code is enough to identify 85% of individuals in the United States. Another presenter added that a tax ID number was a preferred identifier for international students or students who may face immigration issues. Despite this seemingly high coverage, panelists noted troublesome gaps in such data—including that the 15% of students who fall outside of the name, date of birth, and zip code coverage are not a random sample of students. To avoid potential bias, presenters and panelists discussed that the conversation should begin with SSN, taxpayer ID number, and/or passport ID number which could then be converted into a federal student ID.

Opportunities to Reduce Institutional Burden and Leverage Existing Structures

All presenters expressed the need for a federal SLDN to reduce burden on institutions that report data, a goal echoed in the CTA. Panelists agreed that burden reduction could function to increase buy-in and participation and improve data usage.

In discussions of burden reduction, panelists highlighted possible approaches to leveraging existing structures in ways that could streamline the data reporting process and limit the workload for NCES. For example, a panelist noted that IPEDS currently collects data from almost 6,000 institutions individually—an arrangement that requires investment from each institution as well as ED. A presenter with experience in state-level data collection noted that this approach could be streamlined by supporting and expanding the state-level collection of student data. The presenter argued that many states currently collect postsecondary data, doing so independently of—though potentially complementary to—IPEDS reporting and other federal efforts. Building on this idea, presenters and panelists discussed a scenario whereby state-level collection could expand to all states and territories, and NCES could partner with these state or territory systems to achieve the same level of data collection. Presenters and panelists noted that this scenario would require NCES to engage roughly 57 state or territory systems, rather than thousands of institutions reporting independently. For this approach to meet the needs of the SLDN, states would need to close information gaps, including the absence of private institutions, that exist in their data systems.

If state systems could streamline the process of collecting data from institutions and uploading that information on behalf of institutions, this approach would substantially lower the burden on institutions, which would only need to report data to one system. This arrangement would also support state systems, which presenters and panelists expect would continue regardless of a federal SLDN.

Another benefit of encouraging state submission of institution data is that this proposed arrangement could reduce stress on the federal data collection help desk, as it would engage primarily with state systems rather than individual institutions. This would also reduce the outreach and effort required to conduct quality control and follow-up. An additional potential benefit of this state-level approach is that it would engage individuals at all levels—that is, at institutions, states, and the federal government—in data collection efforts, which could in turn increase buy-in broadly within the higher education community. Another proposed option was a hybrid model, in which institutions located in states with robust student-level data systems have their data reported by their state, while other institutions report data to the federal system directly.

Presenters and panelists discussed the design of Model 3 to reduce institutional burden—that is, architecture allowing institutions to submit data in their native format, paired with a codebook that SLDN administrators could use to standardize the data. Panelists noted that this approach would require less work for institutional staff because the staff would not need to make the data meet the requirements of a federal system. An added benefit for the federal SLDN would be a level of transparency made possible by the codebook as well as quality control achieved through confirmation rule sets that would be applied to the data to meet formatting requirements. Panelists agreed that, if possible at scale, this approach could represent a powerful advancement of the data collection methodology. It was also noted, however, that operationalizing this approach at scale would be a challenge.

Data Revisions and Storage

Panelists raised the question to the presenters as to who owns the data, noting that ownership has implications for revisions, storage, and access. Presenters and panelists debated this point, with some panelists noting that institutions should own the data they provide and thus have the capacity to make continuous updates to that data or, at minimum, periodic edits. Other participants noted that student information, like demographics, can change over time and the data collection system should be able to accommodate such changes.

Another perspective offered by panelists was that once ED adjudicates data reported to the SLDN, that information becomes static and is owned by the government. This static condition would be necessary for consumers and researchers who expect that data used for analyses be consistent over time. Some panelists further noted that any updates institutions may want to make would represent small variations that would not affect average aggregate statistics calculated at this scale, and thus would not be necessary to be made in real time.

Panelists observed that a static approach may serve users but could cause issues for those who report data. Inconsistencies that are considered noise at an SLDN scale could have serious implications for data analyzed at an institutional level. For example, one of the presenters of Model 2 pointed out that it is not noise if a university president or provost sees data reported for an institution and asks why it does not match other records. The presenters went on to propose a compromise solution that could allow for updating of data along with a qualifier that could inform data users when information had been updated.

Presenters and panelists noted that the SLDN legislation requires data to be made available to qualified researchers and be suitable for analyses relevant to policymakers, researchers, and students exploring their higher education opportunities. Presenters and panelists discussed options that could achieve this goal, including the use of synthetic data[10] that could be made available to qualified researchers. Panelists also noted that researchers who require access to the data would be subject to additional vetting and approval. Restricted data-use licenses were mentioned as an option, as was the creation of 15–20 secure data centers across the country where licensed users could access data and conduct research in person.

Capacity to Accommodate Additional Data Elements

All models envisioned a platform wherein institutions could access and examine data that they upload to the federal SLDN to create reports or run analyses of internal relevance. Participants considered this feature important for buy-in and posited it as leading to improved data quality.

All models envisioned a platform wherein institutions could access and examine data that they upload to the federal SLDN to create reports or run analyses of internal relevance.

A panelist also noted that the list of variables mandated by the legislation may not address all the information that an institution would find relevant for such analyses. One team of presenters envisioned that the SLDN could promote buy-in and increase engagement by offering the opportunity for institutions to provide additional information in data submissions that only their institution could access. For example, institutions could include identifiers and data elements in the data submission that were relevant to the individual institution but not required for submission. Institutions would then have the ability to analyze the required data, as well as the elective data, on the same platform.

Panelists were asked for their thoughts on whether the models could accommodate additional data and whether they believed institutions would be willing to upload that supplementary information to a federal database. One expert voiced skepticism of the federal government's willingness to support a platform wherein representatives from all institutions could regularly analyze and query data. Another expert noted that the handful of nonrequired variables—for example, high school code, institutionally defined academic units, and other student demographics not captured in federal collections—that would be useful for institutional analyses but not required by the federal government could be useful to other data users. A third panelist noted that the platform should have the ability to adapt in the future and that institutions could potentially pay for computational and analytical access to the system.

NEXT STEPS

While the purpose of the federal SLDN is clearly articulated in the CTA, the form that this data collection would take is still to be determined. Implementing an SLDN would represent a substantial undertaking for both NCES and thousands of postsecondary educational institutions, so thoughtful use of the 4-year transition period is essential. To ensure that the system meets its intended promise, it is important that the designers of the SLDN use this opportunity to advance federal data collection efforts, particularly in terms of increased data use, improved privacy, and reduced burden on data providers. By deliberately engaging experts with a range of perspectives and knowledge, RTI and IHEP hope they have identified potential pain points and solutions related to data submission, data governance, information security, and data usage.

To ensure that the system meets its intended promise, it is important that the designers of the SLDN use this opportunity to advance federal data collection efforts, particularly in terms of increased data use, improved privacy, and reduced burden on data providers.

RTI and IHEP intend to maintain this open dialogue on topics relevant to the proposed federal SLDN. The goal of these forums and any next steps is to give voice to ideas and concepts that could contribute to the design and development of the system and to further support the SLDN's stated goal of reducing burden on institutions and generating improved data that are useful to institutions, researchers, and students.

APPENDIX A:

EXPERT TEAMS AND BRIEF BIOGRAPHIES

TEAM 1

Tod Massa

Policy Analytics Director
State Council of Higher Education for Virginia

Tod Massa started at the State Council of Higher Education for Virginia (SCHEV) in 2001 with nearly a decade of experience in institutional research, policy research and analysis, and data management. As Policy Analytics Director, he oversees the Commonwealth's postsecondary education data system and biennial projections of enrollment and demand; he is one of the architects and leaders of the Virginia Longitudinal Data System (VLDS). He facilitates the monthly VLDS Data Governance Council, manages the support contracts, and work s to expand its membership.

SCHEV has been collecting student-level data on public and nonprofit colleges in the Commonwealth since 1992. These data represent a core strategic information resource for the Commonwealth that Mr. Massa and his team collect, manage, and analyze for higher education policy development. These data also represent one of the cornerstones of VLDS.

Prior to joining SCHEV, Mr. Massa held institutional research leadership positions at two institutions of higher education. From 1991 to 1994, he served as the Institutional Research Coordinator and Academic Policy Analyst at Saint Louis University. Most recently, he served as the Director of Institutional Research and Planning Support at Willamette University in Salem, Oregon. Prior to joining Willamette, he served in the Airmobile Infantry with the U.S. Army and in the U.S. Army Reserve as a platoon sergeant for a medical clearing company, illustrator, and psychological operations team leader.

Mr. Massa studied in the Public Policy Analysis and Planning program at Saint Louis University. He earned a Master of Public Administration from Southern Illinois University Edwardsville. He holds a Bachelor of Arts in studio art from Missouri Southern State College.

Michelle Appel

Assistant Vice President, Institutional Research, Planning, and Assessment
University of Maryland

Michelle Appel is the Assistant Vice President of Institutional Research, Planning, and Assessment, overseeing teams handling policy analysis, institutional and learning outcomes analysis, systems implementation and enrollment projections, and support for campus initiatives through informed decision-making. She has been with the university since 2002, after serving in institutional research positions at Anne Arundel and Carroll Community Colleges.

A past President of the Association for Institutional Research (AIR) Board of Directors, Ms. Appel led the development of AIR's Statement of Ethical Principles, which was approved by the Board in 2019. She frequently presents on the use of data to improve student success and to inform policy. Additionally, she has extensive experience implementing software and has presented on the application of software development techniques to institutional research work. She received the 2010 Distinguished Service Award from the Northeast AIR and the 2006 Marilyn Brown Distinguished Service Award from the Maryland AIR.

Ms. Appel earned her bachelor's degree in Psychology from Bloomsburg University and her master's degree in Human Services Psychology from the University of Maryland, Baltimore County, where she completed the coursework and qualifying examinations for a doctorate in Community and Social Psychology.

TEAM 2

Richard J. Reeves

Associate Vice President
Institutional Research and Effectiveness
University of Delaware

Richard Reeves joined the University of Delaware in January 2019. He works on analytic systems and reporting, data governance, the Cost Study, accreditation, and other work related to measuring and estimating the institution. Previously he worked at NCES where he was the Chief of the Postsecondary Branch and led the IPEDS data collection since 2013. He began his federal service at the U.S. Department of Energy where he was the chief of the Survey Methods Branch of the Energy Information Administration (2010–2013). Prior to his work with the federal government, he was the inaugural Director of Research for the National Student Clearinghouse (2006–2009), a Senior Researcher at Cornell University (2000–2006), and the Director of Enrollment Research and Technology at Johns Hopkins University (1997–2000).

A graduate of Frostburg State University, Mr. Reeves earned a master's degree in Experimental Psychology at Towson University and a master's degree in Statistics at Cornell University. He has given numerous presentations, workshops, and keynote addresses on topics related to enrollment management, student estimation, data system design, time to degree, federal reporting standards, and postsecondary education policy.

Valeria Garcia, Ph.D.

Associate Vice President of Decision Support
Office of Decision Support
University of South Florida

Valeria Garcia is the Associate Vice President of the Office of Decision Support at the University of South Florida in Tampa, Florida, managing a portfolio spanning strategic planning, student and faculty data analytics and reporting, institutional research and effectiveness, performance accountability, and information governance. Throughout her nearly 20-year tenure in higher education, she has built a reputation for genuinely, eagerly, and responsibly advancing the institution's strategic direction and promoting robust efforts around student and faculty success through the use of data, information, analytics, collegiality, and ethical decision-making.

Dr. Garcia serves on a variety of groups locally and nationally, bringing her expertise in the use of data and analytics and policy and planning to address current needs and issues impacting higher education. Nationally, she serves as an officer on the Association of Public and Land-Grant Universities' Commission on Information, Measurement, and Analysis and represents the University of South Florida in various national student success initiatives.

TEAM 3

Brendan Aldrich

Co-Founder
Invoke Learning

Brendan Aldrich is the Founder of Invoke Learning, an education-focused data platforms and artificial intelligence (AI) analytics company for higher education. Over the last decade, he has been responsible for data modernization and innovation initiatives across some of the largest city- and statewide college and university systems in the country to effect transformative change through innovative data architectures that empower data-informed decision-making.

A cross-industry data innovations specialist, Mr. Aldrich has also held roles as the Chief Data Officer for the California State University, Office of the Chancellor; Chief Data Officer of Ivy Tech Community College of Indiana; Executive Director of Enterprise Information Management for City Colleges of Chicago; Director of Business Intelligence Analytics for new media start-up Demand Media; and Director of Business Systems for The Travelers Companies, among others.

In recent years, Mr. Aldrich work has been recognized as a top-three finalist in the global 2021 Snowflake Startup Challenge, in the AWS 2018 City on a Cloud "Dream Big" Award category, Gartner's 2017 Data & Analytics Excellence Award for Self-Service Analytics, Hitachi Vantara's 2017 Excellence Award for Social Impact, and Campus Technology Magazine's 2014 Innovators Award for Administrative Systems.

Pegah Parsi

Chief Privacy Officer
University of California, San Diego

Pegah Parsi is the inaugural Chief Privacy Officer for the University of California, San Diego, campus where she spearheads privacy and data protection efforts for the research, educational, and service enterprise. She is passionate about data ethics and privacy as human rights and civil liberties issues and is an advocate for the idea that privacy requires much more than legal compliance.

Ms. Parsi manages a complex portfolio of privacy initiatives related to employees, students, applicants, alumni, and research participants and provides guidance on privacy laws and regulations, such as General Data Protection Regulation (GDPR), FERPA, Health Insurance Portability and Accountability Act, Personal information Protection Law, California privacy laws, and research privacy / Common Rule. Her day may involve anything from a consult on license plate readers to research involving smart devices to using AI and predictive analytics in support of student success.

Prior to the University of California, San Diego, Ms. Parsi was a privacy manager at Stanford University, focusing on medical studies and international collaborations. She is an attorney and holds an MBA. In her spare time, she advises clients on human rights and asylum matters. She is a Veteran, who, among other things, was the Honor Grad of Army Truck Driver school.

APPENDIX B:

LIST OF FORUM PANELISTS

Lande Ajose
Public Policy Institute of California

Angela Bell
University System of Georgia

Matthew Berry
Kentucky Center for Statistics

Subash D'Souza
California State University

Jacob Jackson
Public Policy Institute of California

Dawn Kenney
Central New Mexico Community College

Michael Le
California State Polytechnic University, Humboldt

Bao Le
Association of Public and Land-Grant Universities

James Parker
Dallas College

Patrick Perry
California Student Aid Commission and
Cradle-to-Career Data System

Jessica Shedd
Tulane University

Sean Simone
Rutgers, The State University of New Jersey

Stephanie Straus
Georgetown University

Sean Tierney
Indiana Commission for Higher Education

Kate Tromble
Data Quality Campaign

Kelia Washington
Data Quality Campaign

Christina Whitfield
State Higher Education Executive Officers Association

APPENDIX C:

Working Paper for Each Student-Level Data Network Model

Team 1

Student-Level Data Network Proposal

Michelle Appel, University of Maryland
Todd Massa, State Council of Higher Education for Virginia

May 2022

This narrative provides a high-level description of a means of collecting student-level data from institutions to support current National Center for Education Statistics (NCES) aggregate Integrated Postsecondary Education Data System (IPEDS) collections. The proposal is grounded in the experience gained both administering a state unit record data collection and as a practitioner of institutional research best practices.

Some of us in the higher education community have advocated or agitated for a national student record system for over two decades. The value of enhanced reporting and analysis capabilities of such a system, even with stringent limitations to protect privacy and against misuse, far outstrip the initial cost and disruption of such an effort. IPEDS has generally met the needs of federal reporting well enough, but it has done so with a general lack of precision and increasing complexity of the surveys. In response to the College Transparency Act's (CTA's) call for a student-level data network (SLDN), this proposal addresses these issues and grounds the system within the protections of CTA and existing federal law.

Others of us, particularly those involved in reporting at higher education institutions, have expressed concern that a national student record system could obfuscate the nuance inherent in student-level records and lead to misinterpretation of individual institutional data. Existing systems such as the National Student Clearinghouse (NSC) provide examples of both the benefits and risks of such systems. Standard reports from the NSC are hard to replicate and can contain errors, particularly around the assignment of first time in college. Detailed reports show the promise of such a system but require considerable data cleaning that is tailored to a specific question. The proposed SLDN addresses these issues by allowing for institutional access and transparency.

This proposal balances the value of a new SLDN while providing transparency and access within the constraints of the CTA. We attempt to provide the flexibility necessary for institutions to accurately report their data with the need for consistency to produce meaningful aggregated reports. We pay close attention to the protections required both by the CTA and by our ethical responsibility to students.

In the sections that follow, the assumptions under which the proposal was developed are outlined, followed by a set of features and special considerations. Each collection is then outlined and tied to specific IPEDS reports. The proposal closes with descriptions of the value added by the system as outlined.

Assumptions

This proposal was developed for a panel exploring potential solutions in advance of final passage of federal law. As such, a variety of assumptions were necessary in the absence of final federal regulation.

- Existing federal privacy law does not change: By law, institutions will not be able to receive student records with additional information (i.e., information not supplied by them), except perhaps in some very limited cases, such as enrollment elsewhere.

- Most states (if any) will NOT end their collections: An added federal collection will need to integrate seamlessly, be designed around a state partnership to perform all submissions by the state or be a hybrid of state and institutional submissions. Most existing state collections can produce all or most IPEDS enrollment and completions surveys and generally at least some portion of the remaining surveys.

- The system will exist in the federal cloud or other system as specified:
    - *"Ensure data privacy and security in accordance with standards and guidelines developed by the National Institute of Standards and Technology, and in accordance with any other Federal law relating to privacy or security, including complying with the requirements of subchapter II of chapter 35 of title 44, United States Code, specifying security categorization under the Federal Information Processing Standards, and security requirements, and setting of National Institute of Standards and Technology security baseline controls at the appropriate level"* (https://www.congress.gov/bill/117th-congress/senate-bill/839/text?r=2&s=1#idd8dcdc97e26841fd87135088655b4b96)

- No matched data from other agencies will be stored as part of the student records:
    - *"Does not result in the creation of a single standing, linked Federal database at the Department that maintains the information reported across other Federal agencies"* (https://www.congress.gov/bill/117th-congress/senate-bill/839/text?r=2&s=1#id24e4777aa9764e0eb1350398ff4c15a9)

- Four- to six-year transition (development through full implementation): It will take time to build all the necessary capacity and the historical set of data needed to produce all IPEDS reports. Therefore, there will be a transition period during which institutions will submit both aggregate and unit record data. Further, it will likely be necessary to submit multiple years of cohort records and completion records for graduation rates and outcome measures. This is not unlike in the mid-1990s when public 4-year institutions had to submit multiple cohorts for the first responses under the Student Right-to-Know Act. The CTA requires that the system shall be developed "not later than 4 years after the date of enactment" and we do not assume that it will be fully populated at that point. Based on traditional practice, we assume the first year of operation will be voluntary with mandatory participation in the following year.

- All Title IV institutions included: This means that the system must be flexible enough to accommodate the limited reporting capacity of very small institutions, particularly those that fulfill a very narrow mission.

- Stringent security practices will be in place: To protect students and to comply with federal law, a strong data governance system will be in place. This will mean that any use of the data not explicitly stated in law and regulation will require considerable scrutiny and will likely not be approved.

- The NSC is not going away—the NSC serves many purposes including transcript services and enrollment verification for a variety of third parties. These services are allowed under FERPA and under the contracts each institution has with the NSC and would continue as there is room for many roles for multiple players.

- Definitions and schema for elements will be found in the Common Education Data Standards.

- As noted in the CTA, institutional needs will be considered as part of the system development. It is our assumption that institutions are both users and providers of data in the system.
    - *"(C) DEVELOPMENT PROCESS.—In developing the postsecondary student data system described in this subsection, the Commissioner shall— (I) focus on the needs of— (I) users of the data system; and (II) entities, including institutions of higher education, reporting to the data system."* (https://www.congress.gov/bill/117th-congress/senate-bill/839/text?r=2&s=1#id5b108036d23a41f4a007e2bff4ca8a17)

- Based on experiences in Virginia and conversations during the time of the IPEDS Unit Record Feasibility study in 2004, we assume a higher-touch model of user support is appropriate. For example, in Virginia, those responsible for data submission know who their SCHEV liaison is and are able to develop a working relationship with that person over time. Further, the liaison develops familiarity with the institution and its data, and a trusting relationship is able to be formed. Further discussion will be in the section titled "State Submissions."

Special Considerations

Information Justice

Data are never objective. Data collections reflect the biases and the interests of those doing the collection. We have an opportunity in designing this SLDN to rethink our traditional models of defining peoples and populations and how we give voice and representation to those who have neither. Yes, classifications of race/ethnicity are built into federal law, but those can be viewed as one of many reporting rubrics. We have an opportunity for thoughtful design that should not be missed. While the bill requires race/ethnicity categories to capture *"all the racial groups specified in the most recent American Community Survey of the Bureau of the Census,"* we can go beyond this if there are other categorizations that are meaningful. Further, gender is left undefined, opening up possibilities for broad discussion.

Fluid Identity Demographics

We live in an era where characteristics such as gender and race/ethnicity are more fluid than ever. While our proposal centers around creation and regular submission of a student identity file, gender and race/ethnicity are not part of that record. We recognize the impermanence of these values and propose instead that there is a linked institution-demography record that carries the "current" demography of the student. These records are only submitted at time of the first enrollment at the institution and subsequent to any changes during a student's enrollment. All records are archived in order to support research into the fluidity of these characteristics. We recognize an inherent risk of maintaining such data, however, experience in Virginia has shown the benefit of this archive in confirming or correcting misidentified individuals and recreating data for a given year.

Features

Required

The following features are "must haves" and flow as the natural outcomes of the assumptions and special considerations. Without these, institutional adoption of the new system would be difficult, and the quality of the data would be compromised.

- Seamless/side-by-side integration with state collections: There are a variety of ways in which the new system could integrate with state collections, but to reduce burden on institutions while maintaining consistency in data reported across levels, the system must account for state collections where they exist.

- Where differences in definitions between state and federal collections exist, states will have to effectively crosswalk to the federal definition or modify their collections. This will more often be a problem where an existing national definition does not exist.

- We assume prior-year revisions will be possible and necessary. Yes, this creates challenges for reproducing prior-year reports, but with adequate tracking and change logs, this is not a tremendous problem.

- Ability for institutions to get aggregate reports on specialized cohorts: The burden for institutions must be offset by the ability for institutions to benefit from the system. One such benefit is the ability of institutions to submit their own cohorts of data (either via a flat file upload or institution-specific fields) to the system and receive aggregate reports about those students. This will also allow institutions to validate those data they did provide (i.e., institutions should be able to replicate their own aggregate information on that group of students).

- Institutional definitions based on system parameters: Under the existing IPEDS collection system, the federal definitions are applied within the institutional context to generate the required reports. For example, the full-time status of graduate students is often based on a blend of course taking, research, and other formal responsibilities that are involved in graduate study, and this is allowed under the IPEDS definition. This type of flexibility must continue, even if some of the underlying components (e.g., number of credit hours) are available within the SLDN.

- Allow form-based entry for small institutions: There are many small and narrowly focused Title IV institutions, particularly among the technical and vocational institutions. To allow ease of adoption for these institutions, the system should allow for the manual entry of student data via a web form.

- Sensitivity to information justice, particularly in the context of fluid demographic characteristics: Unit record data bring with it the responsibility of the system to protect those on whom data are being collected. This is particularly critical with regard to characteristics related to an individual's gender and racial identity. The system needs to collect enough data to allow individuals to express themselves while assuring them that it will not be used to harm them. This means diligence with regard to sharing of any individual data, even (or particularly) among government agencies.
    - *"(D) LIMITATION ON USE BY OTHER FEDERAL AGENCIES.—(i) IN GENERAL.—The Commissioner shall not allow any other Federal agency to use data collected under this subsection for any purpose except—(I) for vetted research and evaluation conducted by the other Federal agency, as described in subparagraph (A)(i); or (II) for a purpose explicitly authorized by this Act."* (https://www.congress.gov/bill/117th-congress/senate-bill/839/text?r=2&s=1#id8e80a576f08f4d438e380ee979073b28)

Desired

The features below would ease burden on institutions and/or provide a "value add" to institutions, thus promoting adoption and even enthusiasm for the new system. These features fulfill the promise of the new system across all levels, not just for external researchers or the federal government.

- Detailed Leaver reports via cohort submission (aggregate): The promise of the SLDN has always been that it will allow richer information about students because it will stitch together data from a variety of sources. Institutional leaders and their membership organizations have embraced the SLDN because they want to learn more about their students. The most meaningful way this can be done is to allow institutions or others with appropriate access to submit cohorts of student identifiers and receive aggregate reports that provide meaningful information about destinations and success. Further, for these and other reports, we think institutions/SHEEOs (State Higher Education Executive Officers) should be able to designate peer groups or use auto-generated peer groups based on broad filtering criteria—national, regional, and sector, for example. Such querying is specifically authorized in the legislation for states and should be available to institutions.

- Data availability comparable to the NSC: The NSC currently provides aggregate reports as well as detailed "directory" data to those who participate in the student tracker system. The NSC allows for institutions to block this transmission for the institution as a whole and for students who have a "FERPA Block" on their records. This system allows institutional and student control over the sharing of information for research purposes while permitting meaningful analysis of unit record data that are considered unprotected (i.e., directory information) under FERPA. However, we recognize that this is not specifically authorized and may not be deemed a proper use of the system.

- Multiple file-format options: Institutions use a variety of systems to maintain information on their students and have varying capabilities to extract that information. Thus the SLDN should be flexible enough to allow data transmission in multiple formats. JSON and XML formats can be highly efficient for ingesting data; however, they are not as universally easily prepared as CSV files. There should be options.

- Submissions by SHEEO offices in place of covered institutions: Because many states already have student record systems, institutions should have the option of allowing the state agencies to submit data for them with an institutional validation either within the state system or after submission.

- Minimal changes to long-standing data structures: Institutions, states, and technology vendors have built capacity based on long-standing data conventions (e.g., census dates, definition of a full-time student). Adoption of a new system would be difficult if both the type of submission (unit record instead of aggregate) and the data structures were to change.

- Planned expansion and modifications to collections: There are some areas in which we can reasonably predict modifications to collections, particularly those already being contemplated within NCES; these should be built into the system. Furthermore, the system should be flexible enough that, given appropriate time and planning, it can expand to accommodate a variety of areas in the future.

- Elimination of institutional involvement in sample surveys: When fully implemented, the SLDN should contain all of the information necessary for NCES to conduct its sample surveys (Beginning Postsecondary Students Longitudinal Study, Baccalaureate and Beyond Longitudinal Study, etc.) without additional involvement of the institutions. This will reduce institutional burden and may result in more accurate contact information than what is available from the institutions.

Institutional Collections

The institutional data collections are more than a collection of data about the institution, they are the core framing of the data collection structure and flow. This collection also defines the edits that are applied to the student data to ensure accuracy and completeness. The institutional collections consist of the standard elements of the IPEDS Registration and Institutional Characteristics collection that define the institutions and go beyond that by defining the award levels of students served; the degrees, certificates, and credentials offered; and each of the programs offered associated with those awards. One of the advantages of this is that it sets expectations of what data to expect, providing a base of information for the structure and data quality checks of subsequent submissions.

Student Data Submissions

Student Identity

The quality of the data and their utility in meeting the required reporting hinges upon the accuracy of the student identities and the ability to match the identities over and over again. One recent example of the power of this is the matching of student loan data with the Social Security Administration to discharge the student loans of disabled individuals. A less positive example is the poor data flow between the National Student Loan Data System (NSLDS) and the various loan servicers. It is our contention that we must recognize the likelihood that these data will be used for other proper and desirable purposes within the U.S. Department of Education (ED). Thus, there should be a process for submitting student identity records using a Social Security Number (SSN) that returns a federal or state student identifier that can be used for all future submissions and always tracks back to the same student with a high degree of confidence.

We propose the creation of a new federal SLDN identifier linked to each student's identity record and other identifiers in use. The structure of this ID would identify the institution initiating the student record and at what level of enrollment at the time of the first submission. When an institution submits a student identity record consisting of legal name, date of birth, SSN / Individual Taxpayer Identification Number / Adoption Taxpayer Identification Number, and other data, the system looks for the student and returns an existing SLDN identifier if one exists or generates a new one if the student is a new entry to the system. Each record returned to the institution is flagged as either new or existing.
In addition to the elements mentioned previously, this collection would include information, such as geographic origin and an indicator of recent high school completion status, that would allow for production of items currently in the IPEDS 12-month enrollment such as net migration of first-time students. We recognize that "geographic origin" as currently conceived in IPEDS applies to first-time college students only and that when we consider the totality of the collection and the value of migration analysis beyond first-time students, geographic origin becomes one of those "nonpermanent demographic elements" discussed below. We also believe questions of origin and migration are of value and interest for all levels of students.

Nonpermanent Demography

Along with the student identity record, there is a student demography record that captures all that is potentially changeable: race, ethnicity, gender, military status, dependency status, number of dependents, first-generation status, marital status, and so on. These are tied to the current institution reporting the student's enrollment and may vary over time. Further, they are not carried over from one institution to another and may differ across institutions in which a student is concurrently enrolled, based solely on what a student reports to each institution.

This collection would gather the information needed to disaggregate data based on race and gender across the numerous IPEDS collections.

Enrollment
Enrollment reporting is in two parts: fall and annual. The fall submission is institution census-based or as of a specified date. The record consists of student level, status (new, continuing, transfer, readmit, dual enrollment, etc.) degree level, major (if chosen), student load, term start and end dates, distribution of enrollment modalities (in person, distance, hybrid), and indicators of enrollment in developmental courses. The fall submission would fulfill the IPEDS Fall Enrollment reporting as well as provide the data necessary to establish cohorts for Graduation Rates and the fall cohorts of Outcome Measures.

The annual enrollment submission expands on these fields to include a record for each term of the institution with an indicator of successful completion of each term and first-time, new transfer, and returning student status. This submission would fulfill the IPEDS 12-month enrollment reporting requirements as well as provide the data necessary to establish the spring cohorts for Outcome Measures.

Student Aid and Charges
This collection would include the actual charges students received for a given term and the aid awarded to them, with details about the source of that aid. This would provide information used for the current IPEDS Student Financial Aid collection.

Completions
This collection includes all degrees and certificates awarded within the year. Elements would include term, level, Classification of Instructional Programs code of majors, and primary modality of study. This collection fulfills the reporting requirement of IPEDS Completions as well as provides the information necessary to generate the outcomes in the IPEDS Graduation Rates and Outcome Measures reports.

Transfer and Dropout Validation
In this collection, institutions submit a list of students who have neither completed nor returned and are provided a list of matching students with a known status. This will provide the necessary validation of the IPEDS Outcome Measures.

Value Add With a New System

The movement toward an SLDN has provided considerable justification for its utility to researchers, institutions, and lawmakers, particularly within the context of degree-granting institutions. We would like to highlight two items that would bring additional value, including to students and their families.

Streamlined Student Loan Forgiveness
The student loan processing and payment system as it now stands relies on borrowers to provide information needed for loan forgiveness and is subject to considerable human error. Disabled borrowers are required to submit paperwork for discharge of their loans, and borrowers who work for employers that qualify for loan forgiveness must provide evidence of that work. That process can be streamlined by using the data available in the SLDN—borrowers' employment and social security records would be available and linked to the borrower. The system could, for instance, identify most qualifying employers via their Employer Identification Number and automatically credit borrowers who work in them with the appropriate years of service. This would reduce the burden on borrowers and lower errors due to manual intervention. This direct use is not specifically authorized by the bill, but recent events have shown this to be possible with the NSLDS and Social Security Administration, and the SLDN can be used to strengthen use and power of the NSLDS.

Ability to Track Students Beyond Degree-Granting Institutions

Students take a variety of paths for postsecondary education. Some start at technical and vocational institutions and then move on to get degrees, while others supplement their degrees with additional technical training. Bringing information about these institutions into the same system as degree-granting institutions will allow for a much richer understanding of the postsecondary environment by stakeholders, institutions, and students and families. It is important that these paths are included so that they are "seen."

Ability of Institutions to Understand Student Paths and Successes

As specified by the CTA, design of the new system will take into account the needs of both users of the data and the institutions that submit data. Acceptance of the new system will be enhanced if those institutions are also considered users of the data and if they benefit from using the system to learn more about their students. Solutions that allow for this could include interactive tools for cohort analysis or institutional use columns and cohort definitions.

We have mentioned cohort tracking as a desired feature of the SLDN, wherein qualified users could submit a cohort of student identifiers from their institution and receive an aggregated report on outcomes. Ideally, we would like to see two options here:

1. The ability to specify the desired aggregations (within privacy and confidentiality) limits, particularly if users are able to add columns to the cohort list for the aggregations.
2. The ability to store and update a limited number of columns within the student record that are for institutional use only to define cohorts and sub cohorts for ongoing use. These could be of particular value when matching with the Internal Revenue Service is conducted allowing for the reporting of outcomes for nonstandard cohorts (assuming large enough numbers to maintain confidentiality).

The bottom line is that institutions will be more supportive and engaged in the system if they can get useful information back that they cannot get through other means. Mandated compliance is one thing; shared engagement and value creates a better system.

State Submissions

Almost every state has at least one system- or state-level student data collection. System collections serve radically different purposes than state or federal collections. State collections, particularly those of coordinating bodies, serve purposes similar to that of IPEDS but on a much different schedule, often needing much greater flexibility of use in responding to legislative interests. For these reasons, we do not believe, nor cannot easily conceive of the possibility of their end or replacement with an SLDN—unless the SLDN is designed to allow for the variance of state data needs and allow direct and regular access to data from state-located institutions by each state entity with defined educational policy interest.

Further, leveraging the existence of these collections, and the people and processes involved, could do two things: reduce/eliminate duplication of effort for the institutions and take advantage of the existing relationships between state/system offices and institutional staff to minimize the need to increase help desk activities to support an SLDN. During discussions of the 2004 IPEDS Unit Record proposal, one of things we heard was, "We know *you* and can trust you. More importantly we know who we are going to talk to when there are problems with edits or our files. We never know who we will talk to at RTI or who we would talk to with a bigger collection." Thus, we recommend a case management approach to create that familiarity between data providers and data collectors.

To do this, we recommend the following steps:

- Recognize SHEEOs in federal law: Basically, make them eligible to receive grants and funding in the same manner as State Education Agencies.
- Technical Assistance Grants to SHEEOs to expand/modify collections: Adding hundreds of new institutions (proprietary and nonprofit) to each of the state collections and creating collections where none exist.
- Building Data Flows from Institutions to State to IPEDS: Create a system that flows from institution to state entity (which may or may not be the SHEEO office) and from there to ED.
- Use state collections as a "local" help desk to minimize the size of a federal help desk.

While submission of student-level data is desirable and creates tremendous efficiencies for research and reporting, it does require significant resources for collection, and the number of potential errors in a submission grows accordingly. Virginia uses a help desk of three people to support collections from 72 institutions. They are backed up by others who also support and maintain the system. This ratio works fairly well, even when an institution submits 30,000 or more student records and it takes time to determine what is wrong with one record or why all records are wrong. As mentioned previously, institutional staff always know who they are working with and are able to develop a relationship that makes these things go more smoothly.

A goal of the bill, as articulated in Section 5 is to reduce the reporting burden for entities that report(ed) into IPEDS. This will not be achieved if the state collections are not included in the design and implementation of the SLDN.

Conclusion and Summary

In this proposal we have outlined what an SLDN would look like within the confines of the CTA. We have paid special attention to the considerations of changing demographic identities and information justice, elements critical in a modern student data system. The proposal is grounded in our experiences and judgment of what is necessary to not only comply with the law as written but to deliver on the promise of the law and engage individuals who work with student data at the institutional, state, and national levels. Such engagement and buy-in will enrich the quality of the data and of the ultimate product.

Team 2

Building a Scalable Student Unit-Level Data Model for Reporting

Richard J. Reeves, University of Delaware
Valeria Garcia, Ph.D., University of South Florida

May 2022

Abstract

All Title IV Institutions in the United States and its territories participate in federal data collections that are required by law. The resulting federal collections have established data standards that can be used for a framework of systems reporting. As a part of the National Center for Education Statistics (NCES), the federal entity tasked with collection of education-related data, the Integrated Postsecondary Education Data System (IPEDS) is a data collection that consists of student data collections and other administrative data collections. Student data collections are fall enrollment, 12-month enrollment, completions, 4-year graduation rate, and 8-year graduation rate. A multitenant system leveraging Infrastructure as a Service (IaaS) and Platform as a Service (PaaS) allows for institutions to load data into the application using established data structures and governance. Files are processed for IDs, examined for errors at the student level and in aggregate, and migrated back into an analytic reporting module for tenants as well as a public reporting model. The anticipated result is a data collection and reporting platform that offers a value proposition for higher education institutions, federal reporting entities, and the public in the form of reinforcement of data and reporting standards, reporting efficiencies, and expanded access to aggregate data for research and analytics.

***Keywords*: Student-level data network, Federal Student Information System, IPEDS, NCES, graduation rates, enrollment, completion, retention**

Executive Summary

Every institution in the United States that participates in the Federal Student Aid Programs (Title IV) submits data to the U.S. Department of Education, National Center for Education Statistics (NCES), as the federal entity for collection of education-related data. Specifically for postsecondary institutions, data are submitted through the Integrated Postsecondary Education Data System (IPEDS) survey collection. The IPEDS collection has components that are student based (e.g., enrollment, completions, graduation rates, student financial aid) and those that are not student based (e.g., human resources, finance, institutional characteristics, admissions). Every institution takes student-level data from its administrative systems and aggregates it based on data definitions established by NCES, federal law, federal regulation, and institutional standards. This paper proposes a model of student record submission to a federal platform that 1) removes the need for aggregation by the institution, 2) enhances the public's ability to conduct research on students, 3) provides additional analytic support to all institutions, and 4) offers a scalable solution.

Historically, information systems have been siloed and built to support singular structures and schemas. Presently systems are able to leverage platforms, architecture, and software as services. There is no longer a need for a singular instance or database, but rather a multitenant system can offer extreme functionality and flexibility. Multitenant systems or multitenancy offer a single instance of a software application (and its underlying database and hardware service) and serves multiple tenants or users. In this case, a tenant is an institution submitting IPEDS (Keyholder) or a system coordinator (Coordinator). In practice this system provides a common platform to upload student-level data and subsequently use the data for analytic reporting.

From an institutional perspective, this platform will allow institutions to upload student record-level data and review the aggregated statistics from this submission in several ways: first, as a single report on metrics and counts for the students uploaded; second, as a set of trend line reports for the institution (e.g., trends in enrollment); third, as a peer comparison tool that provides access to all of the data collected in IPEDS; and fourth, after processing the record-level data from each institution, NCES may provide subsequent enrollment and degree information of former students from the submitting institution, providing additional analytic support to institutions by answering questions such as "Where did our students go?" Or "How many graduated and with what degree?" Providing subsequent enrollment and degree information for students an institution has already educated is not a release of personal information back to the institution that already was in possession of said personal information; rather, it is a release of directory information as described in the Family Education Rights and Privacy Act (FERPA). Directory information includes enrollment status, graduation, and dates of attendance. Not only will this platform lower the burden of submissions, but it will also empower more institutions to use data in IPEDS for analytics in ways they are not equipped to do today.

From a public perspective this platform will allow for a host of advantages. First, it will provide a continuity of measures using the standardized metrics that postsecondary education uses (e.g., fall enrollment and graduation rates). Second, it will provide for a fair and open method for following students through their postsecondary education (something lost in IPEDS). Third, this platform will allow for a development of new statistics without requiring new data to be collected (e.g., presently the 8-year graduation rate cannot be reported by racial/ethnic status because that survey does not include demographics). Fourth, the platform can be used to create a National Student ID that would facilitate anonymized research with other federal statistical agencies through a matching process that can employ encrypted or masked data from each agency (e.g., income, labor, or health). Fifth, the platform will facilitate a more developed and empowered set of sample surveys because students may elect to allow NCES to use their federal education record as background information (possibly enhanced with geographic, labor, health, criminal, and earnings data from other agencies) before asking important questions designed for each specific student type.

Ultimately, a multitenant student information system would benefit institutions by lowering burden, by increasing the analytic capability available to them, and by providing a more complete (and standardized view) of subsequent enrollment and degree information. The public will benefit from a lower burden to data collection that provides a continuity of statistics, enhanced measures, and more flexibility for future research.

Building a Scalable Student-Level Data Network for Reporting

The Integrated Postsecondary Education Data System (IPEDS) is established as the core postsecondary education data collection through the National Center for Education Statistics (NCES), for all institutions that participate in the federal student financial aid programs commonly referred to as Title IV programs. The collection comprises a set of surveys collecting institution-level data (e.g., Institutional Characteristics, Admissions, Finance, Academic Libraries, registration, report mapping, institution identification, IC-Header). The student collections fall into three basic types of information: enrollment, completion, and attributes. This paper proposes a model of student record submission to a federal platform that 1) removes the need for aggregation by the institution, 2) enhances the public's ability to conduct research on students, 3) provides additional analytic support to all institutions, and 4) offers a scalable solution.

Assumptions and Considerations

- **IPEDS**

Institutional and administrative aspects of the IPEDS collection will continue and the metrics established in IPEDS, like fall enrollment or graduation rates, will have value as standardized measures and their importance will remain regardless of the level of detail in the data collection. There is an assumption that readers understand the limitation of aggregate data collections and the rigidity they impose. Institutional framing of cohorts will remain important.

- **Privacy and Security**

The federal government and more specifically the federal statistical agencies have a strong track record of protecting the privacy of the records they collect. For agencies with decades of experience in the protection of census data, labor data, tax data, wage record data, and so on, we can see that the data breaches associated with personal information are limited compared with private-sector actors. This platform should be built with the same care and levels of security and professionalism as other platforms.

- **Platform Architecture**

A student-level data network (SLDN) should be designed using a multitenant application where the management of users are organized around institutional credentials where users may see the data from their institution but not data of other tenants (from other institutions). The platform should adhere to the standardized cloud architecture (e.g., AWS GovCloud), and the specific solution will have some common functionality including a container with a schema (e.g., Redshift or Oracle), an analytic summary tool (e.g., Tableau or Clique), and a user interface and management solution (e.g., Azure).

- **Costs**

The costs associated with the IPEDS collection will increase. However, there may be greater value in the utility of the data for the institution (the institutional value proposition) and ultimately a lower burden estimate for the institutions.

- **Value Proposition**

The value proposition of this platform is multidimensional, spanning higher education institutions of all types, federal reporting entities, and the public in the form of reinforcement and continuity of data and reporting standards already established, ability to utilize data already collected for new analyses and reporting, enhanced reporting efficiencies, and expanded access to aggregate data for research and analytics.

About Student Data in IPEDS

Enrollment data are those data that are used to count students enrolled at each institution. Some institutions use a semester or trimester academic calendar while others use a periodic or on-demand academic calendar. In all cases the fall enrollment is meant to be a single point in time enrollment while the 12-month enrollment represents a count of total students served. The enrollment data layout would satisfy both collections and it is likely easier for institutions to submit the same enrollment file for every term, or for institutions with ongoing terms, a monthly submission makes sense. Data that are aligned with enrollment are those pieces of information specific to the enrollment period like academic load, program(s) of study, credits attempted, credits earned, and financial charges for the term. For many institutions, students may be enrolled and taking classes towards a degree without declaring a program (major) of study.

The concept that enrollment is not necessarily coinciding with a major, or program of study, is an important concept in administrative data collection. That is, often the constructs that policy analysts want to study together are not necessarily required to coexist in practice and thus difficult to study. Often an aggregate collection masks these underlying practices of administrative collection. New data standards will need to be created to address a myriad of assumptions made about student enrollment data that are not absolute.

Completion data are collected at a program and degree level. Because a single person may earn multiple degrees, the idea of completers (unique people finishing a program) or completions (total number of degrees awarded) are two different ideas. As with enrollment data, completion data can offer complexities that inhibit policy analysis (e.g., there is a difference between a dual degree and a double major). Completion data, collected at the program level cannot currently be aligned with enrollment data to draw conclusions on time to degree or pathways to degree. The aggregate nature of the current data collection makes it impossible to truly understand pathways to a degree.

Attribute data are those data about the student that can persist and initiate outside of enrollment and completion data. The cohort a student entered at an institution, gender, race/ethnicity, first-generation status, and so on, can remain in student attributes and be uploaded a single time at matriculation or updated when an attribute changes. Important to the attribute data is the maintenance of a time/date stamp for the records such that they coincide with their first enrollment, most recently. The consideration here is in how the enrollment and degree records should be constructed based on attributes (current, cohort/original) independent of those attributes themselves.

Current Aggregate and Future Record-Level Student Data Collection

Enrollment, completion, and student attribute data (student data) exist at a student level at the institution and are collected and organized for administrative functions related to the enrollment, education, and reporting of the students. IPEDS provides data definitions that are adequate for the assembly and submission of data to the federal government. Student attribute data for the purposes of this paper are established at matriculation. Presently, institutions aggregate record-level student data, assembling it in a set of enrollment or completion data from record-level data into a set of aggregate reports using attributes and then send it to IPEDS. We suggest that the submission of aggregated data is more burdensome than submitting the unit-level data to the U.S. Department of Education (ED), that record-level data are easier to automate, and that it is easier to perform integrity checks on raw data, ensuring consistency. In addition, the aggregated data are more limited for research purposes and more time consuming to assemble. Furthermore, the submission process was designed in the early 2000s and has had regular updates; however, NCES, and more broadly, ED only release data back to institutions 6 months or more after submission.

A future system allows for the submission of data on a regular schedule, spanning enrollment, completion, student financial aid, admissions, metadata related to institutional operations (e.g., term beginning and end dates, degree award dates), and data related to programs offered. Data for enrollment would have a different cadence of submission (e.g., monthly) while student demographics (e.g., cohort type, gender, race/ethnicity, first-generation status) could be submitted annually, and the institution would only need to send a new demographic record if there were a change. Thus, student demographics are only resubmitted if a demographic changes (say annually), while enrollment information can be sent monthly. If a student's demographic information changes, the new information can supplant older information for new reports or remain consistent with the cohort definition. Data standardization on changes in demographic data have not been made federally and will need to be before reporting can be created.

As currently written, the legislation fails to recognize the significance of student financial data and its importance to the U.S. economy and choices made by student outcomes. Institutions can submit student financial aid data that include institutional, state, and federal packaging information as well as key factors of the student aid application. The IPEDS Student Financial Aid survey has this pertinent information, and a more complete set of student financial aid and billing information could be organized in the collection. However, institutions also submit these data to the National Student Loan Data System (NSLDS), and it begs the question if institutions could route the same data sent there to NCES.

Platform Architecture

The institutional tenancy is part of a multitenant platform and infrastructure that works the same way for institutions that currently submit to and use IPEDS. Just as Software as a Service (SaaS) has revolutionized how cloud-based applications can be used, there are Infrastructure as a Service (IaaS) and Platform as a Service (PaaS) technologies that can amplify the efficiency of a data collection and reporting platform. Current examples of a SaaS solution that many readers may have experienced are Office365, Google Workspace, and Dropbox. In each of these cases a user may access the applications using a variety of Internet browsers regardless of what machine the user is connecting with.

A PaaS, as part of its construction, is an analytic module available to the tenants (submitting institutions) whereby they could conduct analysis and build reports using the data sent to NCES for institutional research and reporting. This system as proposed currently would not enlighten those institutions that struggle to report to IPEDS presently; however, there is a possibility that a secondary submission application or middleware could be created to facilitate data submission (e.g., IPEDS in a box). Some of these reports could facilitate checking the data submission in the platform prior to submission, thus reducing total errors. It is these services provided by NCES to institutions that give the value proposition to institutions and facilitate a better institutional research and data use community in postsecondary education.

The legislation indicates that program and institution-level information are in feedback reports. Given that each institution already knows the personal information for its own students, it is not clear why subsequent enrollment and degree information would not be provided to institutions that had enrolled the students previously. Enrollment and degree information are directory information as defined by the Family Educational Rights and Privacy Act. Sharing this information with previously enrolled institutions is not a threat to privacy and is important for an institution's ability to evaluate program and matriculation standards with regard to how they relate to student outcomes.

Consider the following submission process from an institution:

1. The institution generates an enrollment file for the term and uploads it into a platform that

2. generates a set of reports that show the institutional data over time and for that term, and if the reports look accurate to the institutional submitter, then it accepts the data submission and pushes for federal review (this is similar to tax submission processes where a person can review financial data with prior-year context before submitting taxes).

3. The federal review of data is separated into two activities:

    A. The records are migrated into a platform using student personally identifiable information, and a National Student ID (NSID) is assigned to novel students or matched to existing students. The NSID is matched to the submission, and the name, date of birth, Social Security Number (SSN), and other externally identifiable information are removed. The NSID is the only part of the submission that migrates to the analytic system with the educational record. The NSID is matched back to the institution so that an institution may know its student's NSID.

    B. The enrollment, degree, financial, or other student data are loaded into a platform for national reporting and statistical analysis. With the NSID only, an operator can review records and perform data checks but will not know who the individual is. NSIDs for former students matched to a new institution may be passed back for subsequent enrollment and degree attainment reporting and analysis.

4. The external research facilitated by the proposed law would allow for agencies to submit lists of people to the platform to search for an NSID, pass that NSID into the educational record system and return a deidentified data set with educational records and whatever elements were attached to the original records (i.e., salary data), allowing agencies to conduct interdepartmental anonymized research.

It is noteworthy that the taxpayer ID or SSN is a federal ID that has been co-opted by the financial industry. The benefit to corporations and costs to the taxpayer by the private sector using SSNs has been extreme. Efforts should be made to disallow nongovernment agencies using an NSID for noneducational purposes. This paper will not contemplate all privacy provisions in the purely hypothetical (without a law) but the intention is to keep education records with an NSID separate from other identifiable information like name, date of birth, or SSN. Thus after processing and merging, the educational research records are uniquely identified, but not with information useful to the public. The idea that an NSID is only really known by the student, educational institution, and ED should both assuage privacy concerns and cause critical consideration for why a taxpayer ID is used for anything but paying taxes and receiving earned benefits from the government.

Cost

According to USAspending.gov data,[1] the IPEDS contract has an award amount of $82.9 million presently, and that amount encapsulates roughly 9 years or approximately $9 million dollars a year. Considering this amount includes the collection of 12 survey components, over three collection periods, from over 7,000 institutions, that means that the IPEDS collections is running about $425 for each institution for each collection period that includes an average of four survey components ($9 million / 3 / 7,000 = $428). In all of this, we have not included the cost of registration and the creation of the header. This low price includes building reports and supporting a call center, a data center, and data feeds to other products related to IPEDS. Those costs will likely not go away because of a new collection mode. In fact, there may be a higher demand on call centers and report activity in the beginning of a new platform migration. The federal government should consider the value of using commercially available cloud-based SaaS, PaaS, and IaaS for three specific reasons:

1. The economy of scale for the platform
2. Using tools already in the marketplace will minimize training costs for colleges
3. Using commercially available tools makes transitioning to newer, better tools easier in the future

The NSLDS already receives much of the information being proposed here and may serve as a proxy for understanding the cost of a system like this. That is not to say that the NSLDS is a system capable of appropriate national statistical reporting because that is not what it was ever designed to do. However, the data feeds going to the NSLDS either directly or by way of the National Student Clearinghouse demonstrate that institutions can prepare student-level files and send them for all students participating in the federal student aid system with enrollment, program, and degree information. The question is, what is the cost of collecting these data for all students and creating statistical reports.

A similar search of the NSLDS contractor shows that Accenture is a contractor on this task[2] and appears to have spent roughly $747 million on this platform since March of 2015 or 6 years, for an average of just over $124 million dollars a year. The question is not should the NSLDS be used for statistical reporting but should statistical reporting be used to provide data to the NSLDS? That answer is "yes" and would lead to a better aligned, more accurate, less fraudulent student aid system that can be accurately used with other federal statistical agencies while supporting student aid in a meaningful way. A realistic cost estimate for a truly national collection is likely the combination of a large portion of the money needed to conduct an NSLDS collection with the entirety of the IPEDS collection. Therefore, a cost estimate between $30 million and $80 million a year is a reasonable budget depending on the level of support this system provides to federal student aid.

Value Proposition

There are several important parts to the value proposition for this plan. First is the value to institutions in creating a more efficient submission process. Presently, IPEDS estimates the total burden for submitting its forms on average as 52–127 hours depending on experience and institution type. These estimates do not include any data reports or the development of comparison data that many institutions do. Second, the reporting to the National Student Clearinghouse, the NSLDS, and other agencies is redundant, burdensome, and inadequate.[3] There is the opportunity to create a system that facilitates data analysis with submission and leaves open the opportunity to report to other entities and further provide efficiency. Lastly, a platform with analytic capability would facilitate a data use culture at institutions using common measures and more standardized statistics.

Policy and Institutional Implications

The landscape of higher education continues to be transformed, shaped by circumstances of time and external factors—for example, the pandemic/endemic impacting delivery of services, evolving workforce demands informing curricula offerings, and changes to student populations and profile highlighting the value proposition of higher education. As institutions of higher education attempt to be responsive to such demands, regulatory reporting demands persist inherent with limited definitions of metrics and restricted access to meaningful data. An SLDN introduces a path to address these challenges, with the multitenant platform and infrastructure proposed in this paper resonating with higher education practitioners given the resources, time, and effort expended to comply with regulatory reporting. The utility of a submission platform with tiered levels (i.e., submission protocols at the institution level, then state and federal levels) not only presents opportunities for reporting efficiencies but also considers the benefits of leveraging analytics and reducing institutional burden.

There are numerous implications for institutions to consider in an SLDN platform. Many of these have been addressed in related forums[4] facilitated over the past several years but warrant further discussion when considering a platform, namely privacy of student records (benefits of a tiered system), how record changes will be governed and managed (governance), inclusion of metrics not currently included in IPEDS submissions (governance), timeline for institutions to be prepared to report on non-IPEDS related metrics, and the frequency of institution participation.

Further, implications should be considered from multiple lenses, for example, a multicampus institution, state governing bodies, varying size and type of institutions, and consortium of institutions. For a multicampus institution, particularly one with separate IPEDS reporting requirements, there is a substantive positive impact to the institutional burden of reporting for the institution since this platform would remove the redundancies of reporting for multiple campuses and would result in the ability to leverage analytics and improve reporting to address institutional questions and research. State governing bodies could assess their own reporting structures to realize additional efficiencies around reporting burdens and analytics. Systems of institutions of higher education could expend their time and energy maximizing data and analytics to follow students through their postsecondary education, across institution types, in turn providing new insights around student success and informing higher education practitioners of new metrics and statistics of importance to the varying student populations being served.

Case Study: Higher Education in Florida

Higher education in the State of Florida consists of a State University System (SUS) made up of 12 public institutions, a Florida College System (FCS) made up of 28 public institutions, and 30 institutions that make up the Independent Colleges and Universities of Florida. For purposes of this paper, the focus is on public institutions (SUS and FCS). Governed by the Florida Board of Governors and the Florida Department of Education (SUS and FCS respectively), the redundancies and inefficiencies surrounding regulatory reporting is a value proposition for the State of Florida and for the numerous other states with 40+ public institutions of higher education (e.g., California, Texas, New York). An SLDN platform, such as the one discussed in this proposal, promotes the ability to mature data currently being captured by IPEDS to account for evolving higher education institution structures (e.g., systems of institutions, multicampus institutions), maximize efficiencies of resources at the institution level and at the state level (e.g., IPEDS keyholder support, report generation), and address concepts such as "swirl" (spanning both enrollment and completion) within an institution across its campuses, across institutions in a state system, and even across state systems of institutions (e.g., SUS and FCS institutions).

The University of South Florida, one of the 12 SUS institutions, is a good use case having had experience with reporting that would benefit from this proposed multitenant SLDN platform. With over a decade of reporting to IPEDS as a multicampus university system (three campuses with separate IPEDS IDs), including separate reporting responsibilities and campus resources required to prepare data submissions for all IPEDS survey, it cannot be overstated that the reporting burden of preparing data files for aggregate reporting (including documentation) exceeds that of student-level record reporting which is also done by each SUS institution to the Florida Board of Governors. State-level reporting processes and systems introduce additional opportunities and considerations for an SLDN platform, which when addressed would further enhance the value proposition for states and institutions. The same reporting and analytics capabilities presented for this model at the institution and federal instance levels could be realized for state governing bodies in their needs for new and enhanced measures in reporting and flexibility for future research and presents an ability to leverage analytics for use in awareness and advocacy, similar to the intent of NCES digest reports.

From the state perspective, the platform presents an opportunity for scalability, under the same security and privacy scrutiny and with adherence to governance protocols, in leveraging the tiered approach of a lower level (institution level at the student record) and a deidentified higher level (federal level) where the assigned student ID allows states to further the work of supporting the success of students as they swirl within and across university and college systems. As another example of the value proposition of such a platform, within the State of Florida, there also exists a Florida Consortium of Metropolitan Research Universities which includes three of Florida's largest SUS institutions, all with enrollments of over 50,000 students—University of South Florida, University of Central Florida, and Florida International University. The consortium strengthens Florida's talent pipeline through sharing best practices and ideas across institutions as well as implementing scalable solutions for student success and access to economic opportunity for the students served. While the consortium has been successful at working collaboratively to advance its goals, a barrier to greater success has been the ability to share and access data in such a way to support relevant and timely student success research and analyses.

Team 3

Data Lakehouses and Privacy by Design: An Architectural Vision for a Proposed Federal Longitudinal Data Network

Brendan Aldrich, Co-founder, Invoke Learning
Pegah K. Parsi, Chief Privacy Officer, University of California, San Diego

May 2022

Introduction

The proposed College Transparency Act (CTA) is important legislation that is broadly supported by more than 150 higher education organizations and other groups. It will authorize the U.S. Department of Education (ED) to build a federal student-level data network (SLDN) to better track the progression of students through their institutions of higher education and forward into the workforce. A letter from the Postsecondary Data Collaborative and the National Skills Coalition, in cooperation with more than 150 higher education organizations and other groups stated,

> *"This bipartisan, bicameral bill would help students and families, policymakers, institutions, and employers to make informed decisions by providing more complete information about college access, success, costs, and outcomes. This information empowers students and families to make well-informed choices about their education, policymakers, and institutions to craft evidence-based policies to help students succeed, and employers to navigate the talent pipeline they need to grow the economy. Without complete, representative data that counts all students, equity will be out of reach."*

The design of such a system is critical to ensuring that it can accomplish the goals set for it. Within the following paper, we offer an architectural vision to ensure that the resulting system is both fully prepared to meet current needs while, as best as possible at this time, positioned to easily accommodate the unknown and as yet undetermined requirements of the future.

Guiding Principles

Great design is accomplished by building upon a strong foundation. In designing this conceptual architecture, we have identified an array of guiding principles for the federal SLDN. These principles embody the need to accommodate existing systems while planning for the future as well as to ensure that the architecture provides for the level of agility and transparency expected of a system designed to meet the needs of the various constituent groups who may use such a platform now and in the future.

These principles are as follows:

- **Hybrid**
  Should the federal SLDN represent a "consolidated" framework (meaning a single warehouse design representing a normalized view of person data across all agencies) or a "federated" design (which would more dynamically facilitate on-demand queries of data from primary stakeholders)? Ultimately, we believe that a "hybrid" approach which incorporates aspects of both approaches and supports the ability for the system to interact with and accommodate all existing and future system requirements is best.

- **Governance**
  The concept of governance, in this case, represents both data ownership and management as well as data responsibility. It is critical that each institution can safely and securely manage its data within the architectural framework (data ownership and management) and for the Commissioner to clearly define the ethical and appropriate uses and disclosures of data held in the federal SLDN (data responsibility).

- **Privacy by Design / Security by Design**
  It is understood that each stakeholder endeavors to ensure the highest levels of privacy and security appropriate for the data of its constituents. The level of attention that these stakeholders employ is to ensure that the information entrusted to ED is not utilized improperly, exposed to inappropriate use, or used in ways that cause harm to individuals or groups. A conceptual architecture for a federal SLDN must be designed to reflect these same high levels of security and privacy that the students expect and deserve when interacting with public entities.

- **Cloud-First**
  Cloud technologies have continued to mature over the last decade. These technologies have achieved a level of maturity that we would entrust student data operations within such a design. In fact, many public agencies have already adopted a "cloud-first" approach to new architectures. This evolution in design thinking recognizes that cloud technologies are not only stable and secure but offer unprecedented levels of cost efficiency, scalability, and robustness that far outpaces traditional on-premises data center operations.

- **Vendor Agnostic**
  While there are many cloud providers available within the marketplace, each provides a number of vendor-specific tools that lead to a certain level of "lock-in" within that platform. On the other hand, a huge number of open-source tools and technologies that work across all major cloud platforms is also available. This team recommends that, wherever possible and cost-feasible, these open-source technologies be utilized to provide ED with the greatest flexibility in design and development. This approach will also allow the federal SLDN to leverage best-of-breed capabilities of each platform as needed or required and enable a future roadmap to a multicloud architecture.

- **Single Source of Truth**
  It is critical that the federal SLDN reflect a single source of truth for all stakeholders. To this end, we recommend a conceptual architecture that facilitates curated data sets within the platform with full visibility to the current and historical rules applied to the transformations within such a resulting data set. These curated data sets must be managed with a cohesive set of data governance standards and should enable the quick development of new rules or even new curated data sets to address evolving problems and challenges not envisaged at this time. We believe the following conceptual architecture fully addresses and accommodates this approach.

- **Transparency**
  The success of this system is fully dependent on achieving a full and supportable level of transparency as to the data and metrics represented in the resulting insights. In recognition of this, we recommend that transparency be represented as a guiding principle. In this respect, we believe that transparency should be reflected in clear and consistent data rules and definitions, clear notifications of changes that take place within the architecture, and a data catalog that contains information related to all data assets and the state of those assets historically through time.

Privacy by Design

The CTA requires that the system must be "privacy protected" (Section 2.1.1.A). The system will house rich profiles and sensitive information about individuals and their families, information that can be used to make informed decisions and empower students and the public. However, as a default, it should be understood that the federal SLDN is not statistically deidentified or anonymized, even where direct identifiers are removed for research purposes (Section 2.1.5.A.i). Even without the use of Social Security Numbers (SSNs) to track someone over time, the level of detail, coupled with demographic information, is sufficient to make these data identifiable regarding many students.

In addition, if the federal SLDN is intended to support decisions that impact individual students, particularly related to employment, the data must be accurate, and any built-in algorithms, artificial intelligence capabilities, or identifiable reports must be 1) ethical, 2) carefully vetted to avoid unintended discrimination or bias, and 3) clear, explainable, and transparent.

For these reasons, we support the privacy-by-design model, where privacy is an essential component of the core functionality of the federal SLDN. Ann Cavoukian, the primary developer of this approach, explained the concept in "Privacy by Design: The 7 Foundational Principles" this way:

Privacy must be incorporated into networked data systems and technologies, by default. Privacy must become integral to organizational priorities, project objectives, design processes, and planning operations. Privacy must be embedded into every standard, protocol, and process that touches our lives.

We use this approach in our proposal. The following are preliminary privacy considerations for the federal SLDN, although the Commissioner should account for all privacy-by-design needs for a robust program (see, e.g., University of California, San Diego's Guiding Principles for Personal Data). We have identified the ones addressed by our proposal; others require further guidance from legislators and/or ED.

- **Purpose Specification**
  The interests, objectives, and purpose(s), including secondary or incidental purposes, for the federal SLDN should be clear from the outset. The design must follow the specified purposes. The CTA implies, but does not clearly state, the purposes for the system. The Commissioner must identify the specific uses of the system. For example, it is clear that the system will be used by policymakers and institutions to improve financial aid programs and institutional processes. Use for scholarly research purposes is also authorized. It is less clear whether the federal SLDN will be used for any transactional needs by organizations or whether institutions and employers can use the system for predictive analytics. Significantly, if the SLDN is intended to assist employers in managing the talent pipeline, that purpose must be clearly delineated.

  We recommend that the Commissioner consider whether individual students and families will receive specific benefits from the system and whether they will have access to any federal SLDN reports or capabilities. Whether or not any benefits will accrue to individual students and families will have additional design implications.

  The intended purposes will drive whether data and reports need to be fully identified, identifiable, or completely deidentified. The particular purposes will also dictate which disclosure limitation methods (e.g., data perturbation, masking, differential privacy) are most appropriate for various releases and the resources necessary for appropriate deidentification. That will also drive who needs to have access to what data and at what level of granularity.

- **Data Accuracy, Integrity, Quality**
  Depending on the uses/purposes of the federal SLDN, some data types may need to be updated in a just-in-time manner, rather than on a set schedule. If, for example, this is used for transactions or to make decisions about a specific group, then a change in a student's name, gender, or sexual orientation may need to be updated immediately. We should, however, be able to see historically how those data may have changed (i.e., it should not override previous information on that person). Universities should also be able to identify why a change was made; for example, a change may be because an institution is correcting an error as opposed to a change to the data subject's characteristics. Our proposal allows for snapshots and as-needed revisions to institutional data.

- **Access**
  The system must allow for different levels of access, based on need and purpose. For example, certain aggregated reports can be available to the public, whereas identifiable data (including reporting of small cell sizes) might require training and agreement to certain data-handling practices. In some circumstances, the views should not be downloadable but be available as read-only, especially those involving sensitive information, such as sexual orientation or race and ethnicity. Our proposed data lakehouse is designed for varying levels of access based on need and has functionality to process data within the system without the ability to download raw data.

- **Accountability and Oversight**
  The CTA will have a direct impact on data subjects (i.e., students and families). Students must be involved in this dialogue and the design of the federal SLDN. The design of the system must be student-centered, and care must be taken not to view individuals as mere data points.

  We strongly urge that the Commissioner create an oversight group (that includes students and privacy subject matter experts) that reviews and monitors uses of the federal SLDN.

  All access should be logged and monitored. Our proposed system is fully logged and auditable. The oversight group should monitor logs.

- **Some Legal Considerations**
  Design of the system will also depend on compliance needs. Some of the data elements proposed for the system have specific legal requirements at the state level. Use of SSNs, in particular, comes with some legal requirements to inform the individual before use. Other elements (e.g., race and ethnicity) are considered sensitive under some state laws and not others.

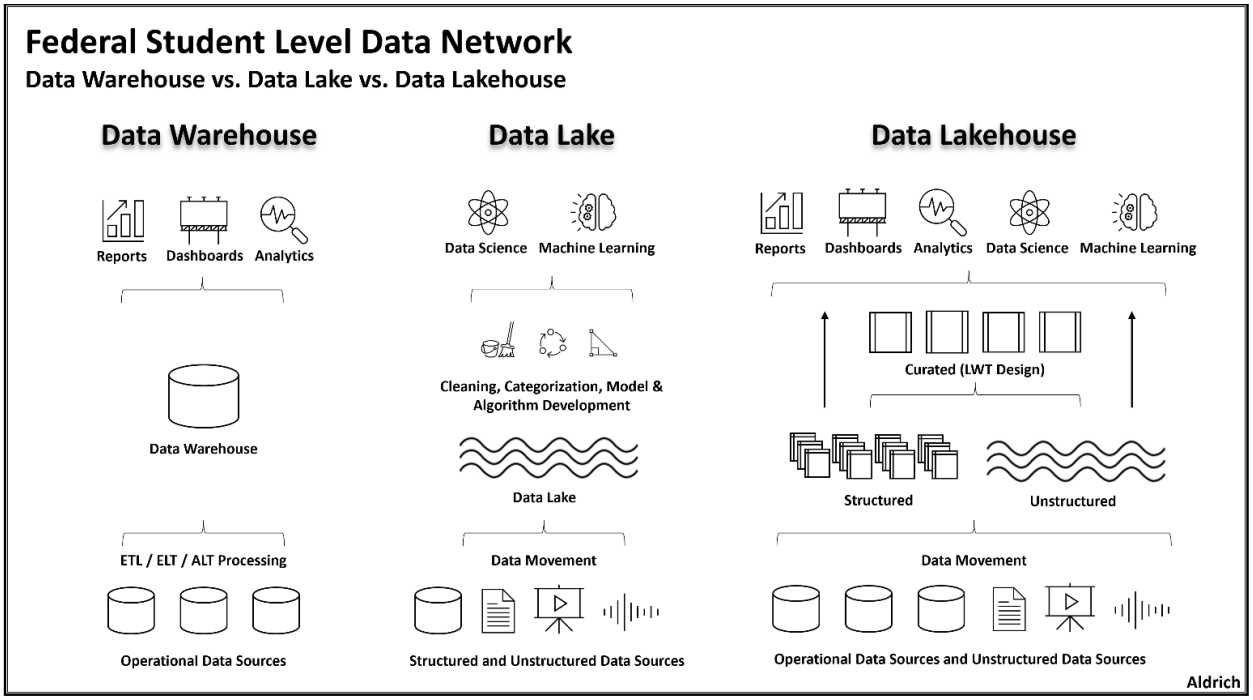
  The federal SLDN will likely contain the data of international students. Because of this, design of the SLDN should account for legal requirements of other jurisdictions, such as those of the European Union's General Data Protection Regulation (GDPR), other GDPR-like legislation in many South American countries, and China's Personal Information Protection Law. There is a risk to universities if they share data of students from those jurisdictions when the federal SLDN is seen as incompatible with those requirements.

- **Other Considerations**
  Universities, and by extension the federal SLDN, are responsible for addressing and responding to data subject requests. The process for addressing such requests will drive the design of the system. What would that process look like, and how would a request be communicated to downstream users? For example, if a university honors an applicant's request to delete the applicant's information, that request must be conveyed to all downstream users of that data, including the federal SLDN. The design must allow for deletions and corrections without reducing the integrity of the remaining data.

Built-in standard reports, functionalities, and analytics capabilities should be available in the platform. The oversight group should assess these for ethics, bias, code reviews, and potential harms prior to release.

This system would be highly valuable for bad actors and should be stress-tested regularly. Security is paramount here.

Data Warehouses, Data Lakes, and Data Lakehouses



The current traditional data warehouse design was first documented by Ralph Kimble and Bill Inmon in the mid-1980s. For more than 35 years, this pattern of design thinking has dominated the conversation around data warehouse design and architecture.

There is no doubt that traditional data warehousing is effective. Organizations around the world utilize data warehouses to manage and optimize key data for reporting and analytics that are critical to their organizations. Unfortunately, it can also be expensive, time consuming, incomplete, and difficult to evolve over time.

In 2010, James Dixon coined the phrase "data lake" and ushered in a new area of technologies that allowed organizations to scale their architectures and capture much more data (volume) of different types (variety) and at faster speeds (velocity). The primary users of these structures are typically data scientists and data analysts using algorithm-based analytics. For their many benefits, these structures also introduced their own challenges around data management, classification, governance, and accessibility.

The latest design thinking in this domain involves an innovative approach to combining the functions of data lakes and data warehouses in a single, cohesive architecture that solves many of the challenges with both while enabling completely new benefits that were previously not feasible or possible. In recent months, this design approach has acquired the name "data lakehouse," and we will both detail and recommend this approach as the design foundation for the federal SLDN.

What Is a "Data Lakehouse"

The data lakehouse was conceived as a method to combine the best elements and capabilities of traditional data warehouses and data lakes into a single architectural technology stack while eliminating many of the challenges inherent in both of these legacy approaches.

As one example, a good data lakehouse design can provide many benefits, including the creation of not one, but two levels of comprehensive historical data stability. This solves a problem with traditional data warehousing where new data elements (if changes are not tracked at the source) must first be added and then trended for a period of time before becoming useful for reporting and analytics. The "degenerated" nature of the Large-Wide-Table (LWT) warehouse design methodology (where the dimensional tables of a traditional star or snowflake schema are folded back into the primary fact tables) also greatly simplifies maintenance, use, and ongoing augmentation.

While the data lakehouse technology approach creates powerful new capabilities, attention MUST be paid to ensuring that the technology is implemented in such a way as to fully realize those capabilities and support the objectives of the federal agency overseeing the initiative. Just as it is possible to design and build bad data warehouses and ineffective data lakes, it is absolutely possible to build a bad data lakehouse.

When done correctly, a data lakehouse will allow organizations to store and processes large volumes of varied data cost effectively—just like the best data lakes—while also providing familiar and effective data structures useful for reporting and analytics—just like the best data warehouses.

Note: For efficiency, we will often refer to a "data lakehouse" as a "lakehouse" over the course of this paper. The terms are intended to be used interchangeably and do not represent separate approaches or technologies.

Foundational Design Considerations

At first glance, the conceptual architecture for a multilevel data lakehouse may look daunting when considered in the context of traditional data warehousing practices that have been in place and practiced across the last 35 years. In actuality, however, several foundational design considerations not only enable this architectural approach but allow us to realize this concept in a manner that is both faster and easier to support than anything using traditional practices. These considerations are addressed below.

Administration at Scale

Automation allows us to administer Institutional Lakehouses, Consortium Lakehouses, Agency Lakehouses, and Other Lakehouses both as individual entities and as groups within a single user interface. This allows us to interface with the entire architecture consistently, reliably, and accurately in the definition and auditing of users, roles, and permissions. In addition, these automation processes can be managed globally, within specific lakehouse groups, or even at the individual lakehouse level.

Administration at scale is a significant benefit of this data lakehouse design approach as management of these activities at scale in more traditional data warehouse and data lake approaches can quickly become fragmented, inconsistent, and inauditable.

Minimization of Data Movement

One of the most common challenges of a "submit and request" approach typically used in a traditional data warehouse is that the data "requested" is extracted from the platform and managed independently by the requesting entity. This data extraction begins to age the moment it is extracted and can quickly become "stale" as the platform continues to be updated with more relevant and historically accurate information.

These "stale" data sets do a disservice to the requesting entity and can create confusion within the industry as the associated insights may be perceived as contradictory to those produced by more current and relevant data.

In a multilevel data lakehouse architecture, minimizing data movement will eliminate the creation of isolated and "stale" data sets across the various lakehouse groupings. As much as is possible, the federal SLDN must empower stakeholders to leverage the platform for common uses to minimize the need for data extraction and creation of these isolated data sets. Minimizing data movement is also highly relevant for security of the data.

The use of "views" or "virtual tables" within the data lakehouse architecture is both highly feasible and desired. In fact, the cloud-first nature of this approach flips the typical dynamics of these structures compared with traditional data warehouses as we can now dynamically scale the compute behind these structures so that they can perform just as quickly as physical tables without creating unnecessary copies of the original data. Reducing the need for stale or shadow data sets is also important for privacy.

Please see the section on record layers for a visual example of data flow between Institutional Lakehouses and the Federal Lakehouse using views and virtual tables to eliminate data movement back and forth between these environments.

Storage Costs and Data Deduplication

In a traditional data environment, storage costs and data duplication become a significant issue as we recommend the submission of all historical data with each file submission by each institution. However, cloud storage costs continually drop to lower and lower levels which helps to mitigate this concern to a large extent. In addition, data deduplication is quickly becoming integrated into a variety of activities within the major cloud platforms to better manage operational function and cost.

The use of data deduplication techniques within the federal SLDN will also reduce data storage and compute costs across these data sets by presenting the ability to use all duplicated data without actually storing every duplicate record.

In this case, a balance must be struck between duplication (faster query execution speeds, but increased storage and compute costs) and deduplication (slower query speeds, but lower storage and compute costs). While the cloud allows us to scale compute to compensate for slower query speeds, this also factors into determining the appropriate balance between useability and cost.

Standards-Based Technology

Using technology standards that are widely adopted will ease and accelerate the speed of adoption by various stakeholders. For example, SQL is a standard database interface language and commonly used across industries. GraphQL, on the other hand, is a newer open-source data query and manipulation language but is not yet familiar across the stakeholders or could be (at least easily) used by those stakeholders to accomplish necessary purposes. This may change over time and, at such time, should GraphQL becomes a common or standard technology, this can be reevaluated as a technology in line with these foundational design considerations.

Standards-Based Connectivity

The most common forms of database connectivity across analytic, reporting, and data science solutions are the Open Database Connectivity and Java Database Connectivity standards. The solution must at least support both connectivity methods and (as much as is possible) be flexible enough to support other broad-based connectivity standards that may become available in the future.

Stated Needs Only or Maximum Future Flexibility?

The proposed design leverages a variety of cloud technologies and approaches to ensure maximum future flexibility while retaining an expected implementation and ongoing support costs that should be lower than if we had designed for even just the stated need utilizing traditional data warehousing technologies and techniques. Some of these cloud capabilities include the following:

- **On-Demand Storage**
  Additional Storage available and allocated upon need as opposed to pre-purchasing storage for future need.

- **Flexible Compute**
  The ability to dynamically apply computing resources on the fly to meet the need as opposed to buying servers at a specific capacity that may or may not meet future need.

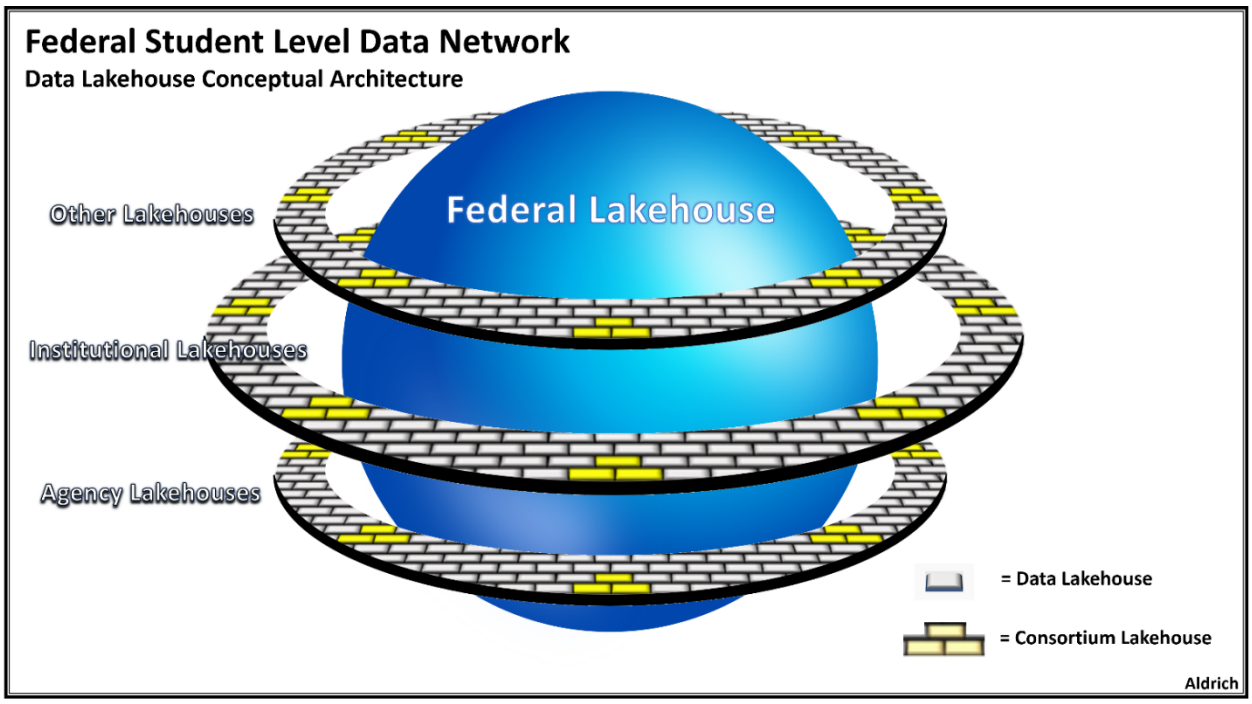
- **Backup and Restore**
  Integrated into cloud at several layers as opposed to a separate application or function applied onto a traditional data infrastructure.

- **Maintenance vs. Use**
  Traditional architectures require expense to maintain and improve the architecture.
  These functions are commoditized in the cloud, which allows us to focus on use as opposed to maintenance.

- **Scalability**
  All architectures have a limit as to how far they can scale. When using the cloud, that limit (while ultimately finite at some level) far exceeds the maximum use of the federal SLDN.

Conceptual Architecture Diagram

The following Data Lakehouse Conceptual Architecture diagram for the federal SLDN embraces a cloud architectural approach that separates "storage" from "compute" and introduces multilevel data lakehouse groups. As a result, the fundamental design recommendation below is referred to as a "multilevel data lakehouse."



The separation of data lakehouse instances into multilevel groups allows us to differentiate data lakehouses by intent and purpose while highlighting the Federal Lakehouse at the center of these connections.

The data storage platform represented within the conceptual architecture consists of exposing data in the form of structures at varying levels of abstraction from their sources. The concept behind the design is to provide security, governance, and standardization while maintaining each institution's ability to control, document, and secure its data individually.

Institutional Lakehouses

The assumption is that each institution has one to many large enterprise databases. Part of this assumption is that the institution databases are structured, relational databases, which have unique schemas and properties such as naming conventions and data types or constraints. The assumption creates the need for the first layer of abstraction, the **Institutional Lakehouse** Layer.

Institutional Lakehouses will ingest institution-specific data into a dedicated, or single-tenant, lakehouse. These lakehouses will be the basic submission of data in a designated format. Submitted files will be natively loaded into a query-able structure with previous submissions, differentiated by a "data date" element. This layer will also include a "conformed" structure that reflects how the native submission has been aligned with agency rule sets. Such rule sets should also be available within this layer as a query-able source.

At this layer the Institutional Lakehouse is intended to be used in federated, agency-managed models ensuring consistency and transparency and transforming the data into formats for large data analytic processing.

- Institutional Lakehouse
  - Raw Files
    - Submission File(s)
    - Codebook(s)
  - Query-able Tables
    - Submission File Table(s)
    - Codebook Table(s)
  - Conformation Rule Set Table(s)
  - Conformed Submission Table(s)
  - Curated Table(s)

Agency Lakehouses

It is expected that the federal SLDN may include data from other federal agencies. In this design, each agency will be provided with its own Agency Lakehouse for the submission and management of data from those sources. The structure will follow the submission process of the Institution and Consortium Lakehouses in that we encourage those agencies to provide full historical data sets in an "as of" format as opposed to incremental additive data sets. This will allow those agencies to correct historical data and provide updates to such historical data that make those time periods more relevant and accurate. Agency Lakehouses are likely to contain (at least) the following:

- Agency Lakehouse
  - Raw Submission Files
  - Query-able Forms of Submission Files

Note: Additional structures may be created as needed dependent on the specific submission and return file needs of a given agency.

Other Lakehouses

There may be a need for specific and/or dedicated lakehouse structures for purposes or groups not identified at this time. This architecture allows for and fully supports this future need with the ability to create additional lakehouses associated with these groups or purposes. File and data structures in these lakehouses should align with data typically made available when feasible, but independent structures may be created and used as needed based on purpose.

- Other Lakehouse
  - TBD

Federal Lakehouse

The Federal Lakehouse represents the final conformed and curated data provided by the institutions. With the right architectural approach, there is no need to physically "move" data from the Institutional Lakehouses or Agency Lakehouses into the Federal Lakehouse. Instead, these data may be sourced from the conformed data sets within each originating lakehouse via one or more views, or virtual tables, within the Federal Lakehouse. These data may then be transformed into an array of physical "curated" tables as needed or desired. Ideally, these curated tables should continue to follow the LWT design.

A view, or virtual version, of these curated tables can then be easily made available to each individual institution as part of its Institutional Lakehouse and will contain only such data from the curated tables as are relevant to that specific institution.

- Federal Lakehouse
  - Query-able Tables
    - Relating to Institutional Lakehouses
  - Submission File Virtual Table
    - Codebook Virtual Table
    - Conformation Rule Set Virtual Table
    - Conformed Submission Table(s)
    - Curated Table(s)
  - Relating to Agency Lakehouses
    - Submission File(s) Virtual Table(s)
    - Return File Table(s)
  - Relating to Other Lakehouses
    - To be determined

Consortium Lakehouses

Within a given set of lakehouses, a "consortium" may be declared. The Consortium Lakehouse will allow institutions to share information among other related institutions, such as colleges within the same district or universities within the same system. The Consortium Lakehouse will contain the data that these specific institutions intend to share with and receive from the federal SLDN. These data sets will need to be well documented and conform to any standardization efforts required to ensure the data can be consumed by other agencies. For each institution participating within a Consortium Lakehouse, the structures include the following:

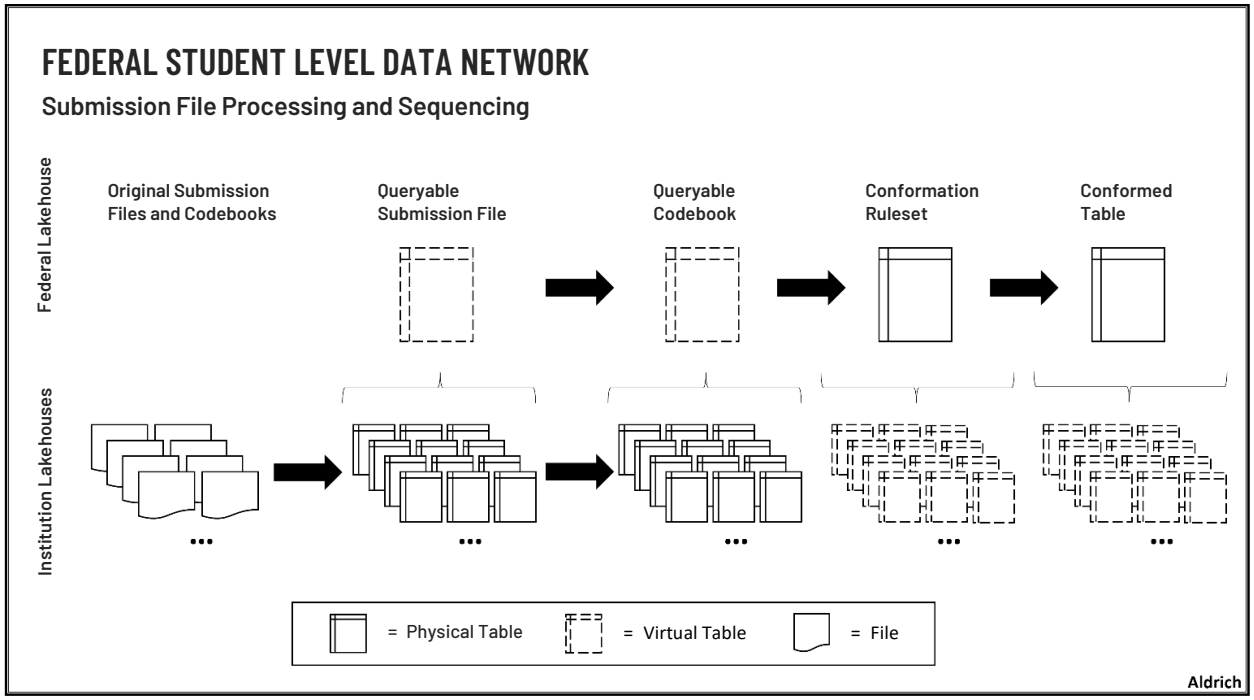
- Consortium Lakehouse
  - Relating to Institutional Lakehouses
    - Submission File Virtual Table
    - Codebook Virtual Table
    - Conformation Rule Set Table
    - Conformed Submission Table(s)
    - Curated Table(s)

In addition, campuses participating in a Consortium Ladkehouse may collaborate with other files and structures specific to the use of the participating campuses.

Record Layers

Use of physical and virtual tables promoting the reduction of data movement is desirable in order to minimize the occurrences of isolated and stale data. The following diagrams show how the record layers between the Institutional Lakehouses and the Federal Lakehouse could interact. For example, when utilizing a single view (or virtual table) to quickly query the data from thousands of individual Institutional Lakehouses submission files (Query-able Submission File, below).

In the opposite direction, a view (or virtual table) within each institution's lakehouse could provide an institution-specific view of data from a conformed table in the Federal Lakehouse (Conformed Table, below).



Submission Files

Submission Files are the initial set of data present within a single data lakehouse. For the purposes of Institutional Lakehouses, this will likely be native data extracted and loaded from their transactional systems but could include other formats, such as processing logs from institutional systems.

Because we recommend the use of data in their native format for a given specification, each college will also upload a "codebook" file. This file will provide textual labels for specific codes for incorporation into conformation rule sets.

Query-able Forms of Submission Files and Codebooks

Submission files are typically of limited use in their original format. As a result, automated processes will be used to load these submission files and codebooks into a query-able table format. This process will be automated and will notify institutions, for example, if values are found in the submission files for which there are no equivalent entries in the codebook. Submission files and codebooks must be aligned for a successful submission by the institution.

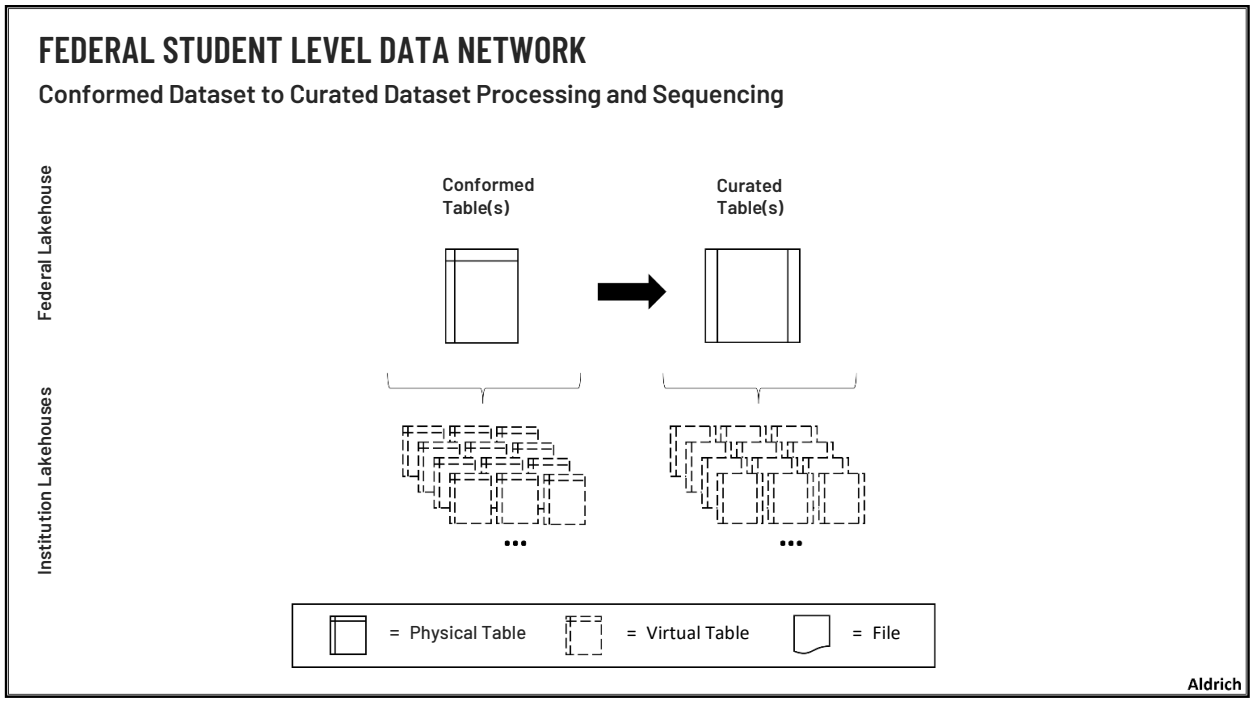
Conformation Rules

The federal agency responsible for management of the Federal Lakehouse will continually review and update the conformation rules to ensure completeness and consistent application of such rules to the submission files provided by institutions. These rule sets will be time and date stamped to ensure transparency as to which rulesets were applied to a given data set across a designated period of time. These rules will govern any necessary or desired "standardization" of data values across institutions.

Because this activity has never been performed at the federal level, this will (at least initially) be more resource intensive to review the data values received by institutions and to build the initial conformation rule sets that will be applied. This activity will subsequently only involve identification and resolution of new values supplied with each submission over time.

Conformed Data Sets

Conformed data sets will be the result of an institution's submission files after the application of the conformation rules. This data set will provide the standardized view of submission files for a given time period and will be the source used for any transformations into the curated data sets.



Curated Data Sets

The product of a data transformation into a form for analysis will be known as a curated data set. A curated data set contains the relevant measures and dimensions for a given level of granularity represented as a single LWT or "Collection Table."

Columnar data platforms, which are prevalent across the major cloud platforms, are significantly different than traditional database structures. Rather than scanning EVERY row and EVERY column of the tables involved in a given query, a columnar data platform still scans EVERY row, but ONLY THE SPECIFIC COLUMNS involved in that query. This allows us to fold the dimensional tables of a traditional star or snowflake schema back into the primary fact tables, creating a unique structure tailored for the modern age: LWTs.

The primary advantage of the LWT design is the simplicity of design and ease of use. For example, rather than ensuring that a series of table joins has been done correctly for a given query, a single query against this table can include data sourced from a variety of systems or locations without a single join—simply a series of "And" statements.

Benefits of the Data Lakehouse Architectural Design

The design of a federal SLDN must provide benefits to every stakeholder and beneficiary involved in the process.

Architectural Benefits

Architectural benefits are described as follows.

- **A Multilevel Usage Group Design**

  A multilevel usage group design allows for gatherings of related data lakehouses into administratively manageable groups. In the current design, the defined groups include the following:

  - Institutional Lakehouses
  - Agency Lakehouses
  - Other Lakehouses
  - Consortium Lakehouses
  - Federal Lakehouse

  Should it be desired, each of these usage groups can be further differentiated for ease of use and administrative purposes. Examples of these subdivisions could include the following:

  - Institution Level: 2-Year, 4-Year, Graduate, etc.
  - Institution Type: Public, Private, etc.
  - Regional Identification: West, Midwest, East, etc.
  - Etc.

- **Simplicity of Analytic Design**
  Use of columnar data structures allows for the degeneration of dimensionality back into the primary fact tables. This approach, referred to as the LWT design, allows us to reduce the complexity of a traditional student data warehouse from 100 to 200 tables requiring complex joins for use down to a core set of comprehensive analytic structure of, say, five tables.

- **Comprehensive Historical Stability**
  Rather than having institutions report a single time period (or "snapshot") that is additive to prior submissions, we propose that the institutions report data for all time periods "as of" a particular submission date. This will provide institutions with the ability to correct historical data as well as to add data of importance in an ongoing fashion—such as retroactively awarded certificates and degrees. Such "as of" submissions are then treated additively, providing the federal SLDN with an ability to track and trend all submitted data elements over time.

- **Incorporation of Public Data Sources**
  This design also incorporates the use of public data extractions from other state, federal, nonprofit, or other entities without disruption to source data or transformed elements. In fact, source data should also be captured in an "as of" form in order to evaluate and consider historical changes to such data sets over time.

  We recommend and encourage the use of Agency Lakehouses to incorporate data from these entities to help provide a more complete picture of students and student needs, which often transcend traditional academic boundaries.

- **Data Submissions in Natural Form**
  In existing approaches to longitudinal data systems, the governing body typically provides instructions to the submitting institutions as to the data and format to be provided. This often requires such institutions to make "judgment calls" as to how they should confirm their data to match the submission standard.

  For example, some such longitudinal systems have required the submission of the student gender to be formatted as "male" or "female." If an institution captures and stores multiple gender options, it is then forced to consider how it should conform the nonbinary gender definitions into the required binary format.

  Some institutions may choose to take all nonbinary students and split them equally among the "male" and "female" categorizations required. Others may use differing approaches, all of which are invisible to a federal longitudinal data system. Even more, the methodology for how each institution categorized its nonbinary students are opaque and may change over time.

  For this reason, we recommend that data values from submitting institutions be provided in as "natural" a state as possible and that any such conformation of that data to consistent values or specific value sets be done at the federal SLDN level using rules that are tracked and made publicly available. This will be accompanied by institutional "codebooks" that provide descriptive labels for the codes provided. This approach leads to standardized views across the country and promotes transparency of the rules used for demographics.

- **Enable Self-Service and Access**
  Rather than the traditional approach of "submit and request" where submissions are made on a required schedule and institutions must "request" data to be returned, we recommend an approach in which each institution has access to its own institutional data lakehouse containing all historically submitted and conformed data sets. This may also include data sets related to other institutions (Consortium Lakehouse). Such consortia may consist, for example, of a multicampus college district or a system of universities within the same state.

Stakeholder Benefits

Because of the wide array of stakeholders and their specific needs for a federal SLDN, the design must be able to provide these benefits for each stakeholder efficiently and in a manner directly useful to the stated purpose. While not all of the following stakeholders are currently specified in the CTA, the following provides a partial list of benefits that we feel this design provides to each group.

- **Students**
  This design encompasses uses for students including, but not limited to, the ability for students and families to easily see what data are associated with their record within the federal SLDN. The federal SLDN can also potentially be used for additional purposes, such as individual degree verifications.

  We also envision other uses for students from this design framework. For example, students could potentially use the federal SLDN for degree verifications or to evaluate themselves against aggregated data relating to similar students or students enrolled in similar institutions. This functionality could be layered upon the Federal Lakehouse or be provided via an "Other Lakehouse" specifically allocated for such a purpose.

  Our proposed design also benefits students and families by incorporating privacy considerations at every stage of design, development, and maintenance.

- **Institutions**
  The design presented within this proposal includes many inherent benefits for institutions that go beyond what may be available from a more traditional structure in a "submit and request" format.

  - Ease of Submission: Because conformation will occur at the level of the Federal Lakehouse, institutions will no longer be required to independently determine their own methods of conforming native values to a required data set. This is somewhat balanced by the need to now submit codebooks that define native values.

  - Transparency of Conformation Rules: An institution may see exactly what rules were applied to its submitted data files at any given time.

  - Incorporation of Retroactive or Historical Changes: Because each submission will contain all history, institutions will be able to have retroactive or historical changes, such as degrees awarded for previous terms, captured and correctly utilized. This allows institutions to correct errors or issues in a previous submission file without the effort of "replacing" a previously integrated data set.

  - Availability of "Curated" Tables: These will provide institutions with the ability to see and use a transformed version of their data as it appears within the Federal Lakehouse.

  - "Live" Data Access: Rather than a "submit and request" format, institutions will be able to interactively access and utilize the data within their institutional data lakehouse and/or the data related to institutions within a declared consortium.

  - Easier Use of Data Across Consortia: With the consortium ability (below), institutions that are part of a district or area of declared relationship will be able to gain visibility to and make better use of the data across these institutions.

- Data Science and Future Use: The nature of the Institutional Lakehouse structure lends itself to additional use by the institution, such as machine learning, data science, and other future uses.

- **Federal Agencies and Policymakers**

  - Incorporation of Retroactive or Historical Changes: Because each submission will contain all history, institutions will be able to have retroactive or historical changes, such as degrees awarded for previous terms, captured and correctly utilized. This allows institutions to correct errors or issues in a previous submission file without the effort of "replacing" a previously integrated data set.

  - "Live" Data Access: Rather than a "submit and request" format, institutions will be able to interactively access and utilize the data within their institutional data lakehouse and/or the data related to institutions within a declared consortium.

- **U.S. Department of Education**

  - All Listed Benefits: We believe this design provides ED with all the benefits that will be seen by individual institutions, consortia, and agencies

- **Consortia**

  - Single Source of Truth: Many colleges work together, either as part of a designated entity (such as multiple colleges in a district or multiple universities within a system) or for other purposes (such as research or federations). It is possible that this collaborative approach may also be necessary or desired by organizations within the "Agency" or "Other" lakehouse groups as well. By facilitating the ability for these institutions to work collaboratively with the data provided by the individual institutions and as conformed and curated within the federal SLDN, we promote transparency and empower those consortia to engage with the federal SLDN more effectively

- **Researchers**

  - Data for Good: The proposed federal SLDN can serve as a powerful tool for education and public policy researchers. In addition to built-in reports and analytic capabilities, the platform will need to provide statistically deidentified data sets appropriate for these types of research purposes.

- **Employers**

  - Future-Forward Design: The CTA is silent on whether employers are intended to benefit from the SLDN. Some interpretations of the text, however, may allow the SLDN to be used by employers, for both individual recruiting and hiring purposes and broader strategic planning. If employers will have access to the SLDN, our proposed ethics oversight group would develop rules to allow for such access.

- **Others**

  - Flexibility in Design: The design proposed within this paper is flexible enough to easily account for specific stakeholders and purposes that are not currently described within the CTA.

Comprehensive Data Governance, Security and Reporting

Comprehensive Data Governance
Challenges with data management cannot be overstated. Issues like poor data quality cost real money, and the federal SLDN process efficiency is negatively impacted by poor data governance. The full potential benefits of the federal SLDN may not be realized if there is not well-documented and understood data governance that is transparent to all stakeholders. Comprehensive Data Governance seeks to remediate issues through guided information management decision-making, ensure information is consistently defined and well understood, increase the use and trust of data as an asset, improve consistency of projects across the organization, and ensure regulatory compliance and eliminate data risks. Enterprise Data Governance includes but is not limited to the following:

- **Master Data Management**
- **Metadata Management**
- **Data Lineage**
- **Data Quality Management**
- **Data Catalog / Data Dictionary**
- **Data Privacy, Security, and Compliance**

As a design consideration, we recommend that all information, data, rules, and documentation (including change tracking over time) related to these areas be fully available in an easily referenced and (ideally) query-able structure that is made available and relevant to all stakeholders.

Information Security and Security by Design

Information Security (Authentication/Authorization/Audit/Encryption/Monitoring/Logging)
Our proposed SLDN design relies on a secure technical foundation, and we recommend that the Commissioner and oversight group develop security protocols for administration of the SLDN.

The main differentiating factor between Privacy and Security, as conceived within this paper, is that Privacy relates to protecting the individual while Security is about protecting all of the data within the federal SLDN.

Much like our proposal around privacy by design, we recommend an approach that incorporates information security—especially as it relates to Confidentiality, Integrity, and Availability—as part of the design itself. Something we could refer to as "security by design."

Cloud providers have a vested interest in security including experts and guidelines to be followed, as do many other higher education groups who make recommendations. Because of its nature, approaches to effective security are constantly changing.

We therefore recommend that such a security by design be implemented as part of the original implementation and evolved as security standards evolve and change. We feel that it would not be prudent to document or propose a particular security by design scheme at this time as it could very well be different or outdated by the time of implementation.

Endnotes:

[1] RTI thanks Jack Stoetzel for his contributions in writing this paper.

[2] RTI International is an independent, nonprofit research institute that works to advance equitable, high-quality learning for people of all ages. RTI aims to inform public policy, expand opportunities, and drive better educational outcomes for children, youth, and adults. RTI conducts several postsecondary education data collections on behalf of the National Center for Education Statistics (NCES). These surveys include the Integrated Postsecondary Education Data System (IPEDS), National Postsecondary Student Aid Study (NPSAS), Baccalaureate and Beyond Longitudinal Study (B&B), and Beginning Postsecondary Students Longitudinal Study (BPS). RTI is acting independently of NCES for this effort.

[3] Dunlop Velez, E., Pretlow, J., & Roberson, A. J. (2020, August). Implementing a federal student-level data network: Advice from experts. RTI International; Institute for Higher Education Policy. https://www.ihep.org/publication/implementing-a-federal-student-level-data-network-advice-from-experts.

[4] Pretlow, J., Dunlop Velez, E., & Roberson, A. J. (2021, January). Implementing a federal student-level data network (part II): Insights from institutional representatives. RTI International; Institute for Higher Education Policy. https://www.ihep.org/publication/implementing-a-student-level-data-network-part-ii-insights-from-institutional-representatives.

[5] Isaac, J., Pretlow, J., Cheng, D., & Roberson, A. J. (2022, January). Implementing a federal student-level data network (part III): Insights from financial aid experts. RTI International; Institute for Higher Education Policy. https://www.ihep.org/publication/sldn-part-iii-financial-aid.

[6] IPEDS Data Explorer, Table 1. Number and percentage distribution of students enrolled at Title IV institutions, by control of institution, student level, level of institution, enrollment status, and other selected characteristics: United States, fall 2020 (Integrated Postsecondary Education Data System) and Table 1. Number and percentage distribution of Title IV institutions, by control of institution, level of institution, and region: United States and other U.S. jurisdictions, academic year 2021–22 (Integrated Postsecondary Education Data System).

[7] The views expressed in the working papers reflect the opinions of the authors and do not necessarily reflect the views of RTI and IHEP.

[8] Under FERPA, schools may disclose, without consent, directory information such as a student's name, address, telephone number, date and place of birth, honors and awards, and dates of attendance.

[9] A data lake is a central storage repository, used to store a large amount of raw data in their native format. The data may be structured, semistructured, or unstructured, which means data can be kept in a flexible format for future use.

[10] Synthetic data are data generated not from actual people or observations; rather, they are artificially produced to mimic actual data. Such data can be used for research purposes while protecting individual privacy.

RTI.ORG

IHEP.ORG