POSTSEC DATA

Ken Ambrose
Senior Advisor
CDO Council, Office of Shared Solutions and Performance Improvement
General Services Administration
1800 F Street, NW
Washington, DC 20405

Dear Mr. Ambrose:

The undersigned 12 members of the Postsecondary Data Collaborative (PostsecData) submit this letter in response to the Chief Data Officer (CDO) Council's Request for Information (RFI).[i] PostsecData is a nonpartisan coalition of organizations committed to the use of high-quality postsecondary data to improve student success and advance educational equity.

Based on our postsecondary data policy research and expertise and interactions with postsecondary stakeholders who are tackling many of these issues, this letter provides answers to the following questions from the CDO Council listed in *Section 4: Data Sharing* and *Section 5: Value and Maturity*.

- What are effective ways for Federal programs to share programmatic data in ways that protect the privacy of individuals and organizations?
- What are the premier examples of public or private sector entities that aggregate, integrate, and share information?
- What are the meaningful approaches to defining the value of government data?

Federal data is most valuable when it is robust, accessible, and protected; PostsecData hopes that the following examples and models can help guide the CDO Council as it aims to improve data practices and access to government data assets.

Considerations for sharing programmatic data while protecting the privacy of individuals and organizations

De-identifying data is one of the most important steps toward ensuring information security, confidentiality, and data integrity, while maximizing data availability and transparency. In many cases, agencies working with personally identifiable data are deliberately vague about how they de-identify the data to safeguard against attempts to reverse the process. Statistical disclosure limitation can be achieved through methods including rounding,[ii] aggregating,[iii] and suppressing[iv] results to obscure unique observations. Additionally, some agencies add in statistical "noise" by changing ages, races, or incomes by a small amount, known as differential privacy.[v] These methods can be applied one-at-a-time or in conjunction with one another. For instance, the College Scorecard, produced by the Department of Education, suppresses the results for any calculation in which the denominator is less than 30 individuals, and also adds in statistical noise to ensure student-data privacy.[vi]

There are some well-established resources designating essential security controls to protect privacy among identifiable information. Perhaps the most stringent guide is the National Institute for Standards and Technology (NIST) Special Publication 800-53 (Rev. 5), Security and Privacy Controls for Federal Information Systems and Organizations.[vii] This document outlines policies, practices, and procedures for ensuring data remains anonymized when in the public domain and secured for authorized use only when needed. These standards and guidelines represent the gold standard, referenced in federal student data legislative proposals including the College Transparency Act.[viii]

Premier examples of entities that aggregate, integrate, and share information.
As highlighted in PostsecData's 2019 release, *Postsecondary Data Infrastructure: What is Possible Today* (authored by Georgetown University's Amy O'Hara, Director of the Federal Statistical Research Data Center), federal and state agencies have developed systems to aggregate, integrate, and share information securely, including information that is sensitive or includes personal identifiers.[ix] Note that several of these examples are outside of the postsecondary space, illustrating that these practices can be useful in all arenas of government, and at federal, state, and local levels.

1) The Centers for Medicare and Medicaid Services (CMS) has implemented a virtual research data center to produce merged extracts of administrative data based on an analyst's needs, including personal identifiers when necessary.[x] Through this system, researchers can use their own laptop to log into a secure environment from which no data leaves.
2) The National Center for Health Statistics (NCHS) offers detailed demographics through remote access to a proprietary virtual research data center. For the analysis of especially sensitive data (e.g., genetic, detailed geography, exact dates, linked files) they have produced physical RDCs.[xi] NCHS maintains a data linkage unit that can match files for analyses, providing integrated data for approved users in the RDC.
3) The Department of Education's College Scorecard is the product of individual-level data drawn from several sources, de-identified and aggregated. Developed from a partnership between the National Center for Education Statistics, the Federal Student Aid administration, and the U.S. Department of Treasury, the Scorecard links students to earnings and student loan repayment outcomes, aggregating them at the institution- and the program-level.
4) Census Bureau has developed the Postsecondary Employment Outcomes (PSEO) program in partnership with 17 states across the nation.[xii] Participating institutions within these states securely provide demographic and completion data to Census Bureau, as well as transcript data. Then, the Bureau matches students with post-graduation economic outcomes (e.g., earnings, employment status), enabling researchers to ascertain how well different programs are serving different types of students. The University of Texas System has taken this partnership a step further by also linking data with the Texas Higher Education Coordinating Board, Texas Workforce Commission, and the National Student

Clearinghouse to produce SeekUT – an interactive dashboard showing aggregated workforce outcomes by program at all University of Texas institutions.[xiii]

Meaningful approaches to define the value of government data

Because it can be used for policymaking, institutional improvement, and for consumer decision-making, postsecondary data has a broad variety of audiences, each finding different components valuable. However, for all of these audiences, government data is most valuable when it is comprehensive, accessible, and contextualized.

First, to be most valuable, postsecondary data should count all students and all outcomes. However, despite improvements in recent years, available postsecondary data is incomplete and not representative of all students and pathways. For instance, the primary dataset revealing post-college earnings outcomes, the College Scorecard, does not disaggregate outcomes by completion status,[xiv] masking the considerable earnings gaps between those who complete their degree and those who did not and producing metrics that are hard for institutional leaders or students to interpret. Collecting and publishing these data can help researchers (and students) to understand the impact of non-completion, especially among students with debt and no degree. Further, the Scorecard only includes earnings information for students who receive federal financial aid because of statutory data collection limitations; these missing data are particularly troublesome for institutions where larger portions of the student body do not receive federal aid, calling into question how representative the data are.

As a second example, recently published data on outcome measures expanded the field's understanding of progression and completion for students who fall outside of the much smaller first-time, full-time cohort captured in the Graduation Rate (GR) survey. Unfortunately, while it disaggregates by attendance intensity, enrollment status, and Pell Grant receipt, OM falls short by not disaggregating by other key student characteristics—like race/ethnicity, gender, and age—necessary to highlight inequities. While improvements are being made in this arena, robust, complete, and disaggregated data would aid institutions and policymakers in identifying the unique barriers and solutions for overcoming these barriers among underrepresented student groups.

Finally, data is at its most valuable when it is publicly accessible and easily comprehensible and contextualized for a broad variety of audiences. For instance, the consumer-facing College Scorecard post-college earnings measures historically included contextual information such as a percentage of students earning more than a high school graduate, as well as national median earnings. By including basic benchmarks, students can easily visualize whether the return on their investment is above average, or at the very least, whether the selected institution prepares them to earn a reasonable wage. Federal data efforts should always consider ways to publish and package data to ensure maximum accessibility and understandability, which will make data use more efficient and effective among all audiences.

***

When developing federal data policy, it is essential to balance the need for useful data to maximize efficiency and inform decision-making against maintaining robust privacy protections to maintain public confidence in data collection and use. We commend the CDO Council's interest in seeking expert input on data sharing and privacy and using it to inform future efforts. If you have any questions about the content of this letter, please contact Mamie Voight, Interim President at the Institute for Higher Education Policy (mvoight@ihep.org or 202-587-4967).

Sincerely,

Achieve Atlanta
Advance CTE
Larry Good, Corporation for a Skilled Workforce
Higher Learning Advocates
Georgetown University Center for Education and the Workforce
Institute for Higher Education Policy
The Institute for College Access and Success
Michael Tamasi, AccuRounds
National Association for College Admission Counseling
National Skills Coalition
Stephen Crawford, George Washington University
uAspire

[i] Federal Register. (2021). Request for information on behalf of the Federal Chief Data Officers Council. Retrieved from: https://www.federalregister.gov/documents/2021/10/14/2021-22267/office-of-shared-solutions-and-performance-improvement-osspi-chief-data-officers-council-cdo-request.

[ii] Cologne, J., Grant, E. J., Nakashima, E., Chen, Y., Funamoto, S., & Katayama, H. (2012). Protecting privacy of shared epidemiologic data without compromising analysis potential. Journal of environmental and public health, 2012, 421989. https://doi.org/10.1155/2012/421989

[iii] Singh, K., & Batten, L. (2017). Aggregating privatized medical data for secure querying applications. Future Generation Computer Systems, 72, 250–263. https://doi.org/10.1016/j.future.2016.11.028

[iv] Sadler, C. (2020). Protecting privacy in data releases. Retrieved from: https://www.newamerica.org/oti/reports/primer-disclosure-limitation/.

[v] Ibid.

[vi] College Scorecard: Data documentation. (2020). Retrieved from: https://collegescorecard.ed.gov/assets/FullDataDocumentation.pdf. f

[vii] Joint Task Force Interagency Working Group. (2020). Security and Privacy Controls for Information Systems and Organizations (Revision 5). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-53r5

[viii] Text - S.839 - 117th Congress (2021-2022): College Transparency Act. (2021, March 18). https://www.congress.gov/bill/117th-congress/senate-bill/839/text

[ix] O'Hara, A. (2019). Postsecondary data infrastructure: What is possible today. Retrieved from: https://www.ihep.org/wp-content/uploads/2019/06/uploads_docs_pubs_ihep_privacy_brief_data_sharing_v2.pdf

[x] Center for Medicare and Medicaid Services. (2021). Virtual Research Data Center. Retrieved from: https://resdac.org/cms-virtual-research-data-center-vrdc.

[xi] National Center for Health Statistics. (2021). Research Data Center (RDC). Retrieved from: https://www.cdc.gov/rdc/index.htm.

[xii] United States Census Bureau. (2021). Post-secondary employment outcomes (PSEO). Retrieved from: https://lehd.ces.census.gov/data/pseo_experimental.html.

[xiii] University of Texas System. (2021). SeekUT. Retrieved from: https://seekut.utsystem.edu/

[xiv] College Scorecard publishes institution-level outcomes using a combined completer and non-completer cohort. For program level data, Scorecard only publishes outcomes for completers, with no published non-completer comparison group.