A Framework for a Multi-State Human Capital Development Data System

Brian T. Prescott
Director of Policy Research
Western Interstate Commission for Higher Education

and

Peter Ewell
Vice President
National Center for Higher Education Management Systems

The rise of a globalized knowledge economy requires us to understand the distribution of skills and abilities in our population. It is no longer sufficient to know how many resources are devoted to the development of our nation's human capital. Today, we also must be able to demonstrate and understand the outcomes of our educational processes.

This growing need has energized interest in building longitudinal data systems capable of following individual students throughout their educational careers. Heightened by the abortive attempt to create a federal student unit data system and three rounds of statewide longitudinal data systems (SLDS) federal grants, the pace has accelerated dramatically with the inclusion of a $250 million funding set-aside for data systems and the required data system assurance by states to access State Fiscal Stabilization Funds in the American Reinvestment and Recovery Act (ARRA). Additionally, efforts are underway to help guide this development, including those being undertaken or funded by the Data Quality Campaign (DQC), the National Center for Higher Education Management Systems (NCHEMS), the National Student Clearinghouse (NSC), the Bill and Melinda Gates Foundation, and the Western Interstate Commission for Higher Education (WICHE).

Despite growing commitment and funding, significant obstacles persist. The intensity of simultaneous activities in this arena may result in efforts that are hurried and uncoordinated, with states independently designing and implementing their own systems. An unfortunate end result may be a patchwork of systems that cannot be easily aligned within a state or across state borders. One example is the lack of coordination nationally with the assignment of unique student identifiers – one of the cornerstones of the database development framework advocated by the DQC. These numbers – critical to linking records for longitudinal tracking – are being put into place state by state, with different structures and attributes, despite the fact that a substantial number of students will cross state lines in the course of their careers. Excessively rapid and uncoordinated database development can and will have unforeseen negative consequences.

This paper presents a framework for how a multi-sector, multi-state data resource might be designed and governed. It is based on discussions and ongoing initiatives across several WICHE states, especially an effort involving the states of Washington, Oregon, Idaho, and Hawaii, to develop a prototype multi-state data exchange (Figures later in this document will refer to these four states for this reason). This "human capital development data system" must be developed to answer "master" policy questions that benefit each of the principal state stakeholders – the K-12 education system, the postsecondary system, and labor/workforce development system – both for accountability purposes and to inform improvements in policy and practice. This requires a review focused on what specific data elements to include in such a system and how to organize them. A particularly prickly

WICHE
Western Interstate Commission for Higher Education

issue facing those who are developing these systems is how to create a workable governance structure and assure that the system is used effectively in ways that ensure security and privacy.

As states struggle to manage scarce resources, accountability in public education increasingly will focus on ensuring that students have acquired the knowledge and skills they need to be competitive in a global economy. There is growing awareness that accountability systems and the databases that support them need not be narrowly focused on compliance but also can be designed to provide information about performance and incentives that lead to improvements in educational outcomes.

Many important, broad-based policy questions can be answered using existing data sources. But to fully reorient our focus on educational outcomes and to disaggregate data to better target interventions, we need more detailed unit-record data that documents educational participation and experiences in both K-12 and postsecondary education, as well as participation in the workforce. Recognizing the shortfalls of our existing capacity to produce policy-relevant information helps clarify the need for more complete data.

Our current array of state accountability indicators in both educational sectors are products of a prior technological environment, in which only aggregate measures applied to individual educational units could be calculated. It was impossible to create a data system that could follow individual students across an entire state or the nation at that time. Instead schools, school districts, and postsecondary institutions became the units of analysis and accountability. The result was a set of cross-sectional views of student progress that do not account for all students, particularly those most at risk. For example, current graduation rates in postsecondary education are based on first-time, full-time students – who constitute a minority of those enrolled at many colleges and universities – and do not disaggregate results by important population characteristics.

To counter this weakness, the federal government has funded an array of longitudinal studies, which go a long way toward proving the utility and power of longitudinal data. In the past, studies that relied on such data have provided useful markers of educational effectiveness and have helped inform some changes in federal policy. Yet these datasets have severe limitations. First, despite the federal government's best efforts, the data are often dated, due to lengthy collection and cleaning processes. Cross-sectional enrollment data drawn from Integrated Postsecondary Education Data System (IPEDS), for instance, generally lags at least two years; and tracking studies based on national longitudinal datasets have been as much as six years out of date by the time they are published. Second, since most educational policies (for example, financing and financial aid policy and policies governing transfer of credit) are promulgated at the state level, a relatively small

dataset representative only at the national level is not of much help in investigating the outcomes associated with individual state policies.

Since most existing data systems are specific to one segment of the educational pipeline, even the best indicators are typically snapshots taken from the inside of a silo. The resulting measures of success have thus focused on processes rather than outcomes, and they are extremely limited in what they can tell us about our success in creating the human capital we need. For instance, a high school might be proud of the percentage of its 9th graders who graduate on time. But if large numbers of those students subsequently fail to enter and complete college, should the high school be eager to declare victory? And what happens to these graduates after their educational experiences are concluded? We don't know much, for example, about what happens to students after they leave our colleges and universities (or schools, especially when they do not move on to college) with respect to their mobility and experiences in the labor market. Do they leave the state? What industries are they working in?

A more effective data system for accountability and policy and practice improvements could provide answers to such questions. Integrated to enable large-scale longitudinal analyses to support state educational and workforce development policy, student or individual unit-record data, linked together across K-12 education, postsecondary education, and the workforce, comprise what we call a human capital development data system (HCDDS). An HCDDS should be able capable of:

- Tracking the stock and flow of the skills and abilities (represented by education and training) of various populations within a given state.
- Examining the gaps in educational attainment between population groups, based on demography and socio-economic status.
- Incorporating information from multiple states, given the mobility of the U.S. population and the fact that many population centers are located on state boundaries.

Reorienting accountability arrangements around longitudinal data acknowledges that developing productive citizens is a core goal of all levels of education. While workforce development is not the only goal of education, it is the one that is the most systematically measureable – through existing data collection activities undertaken by state and federal agencies responsible for labor market information. Focusing on workforce development also acknowledges that students and families throughout the nation typically say that their principal reason for seeking a college education is to improve their employment prospects. Finally, this focus recognizes that many of the other benefits of education – including civic engagement, volunteerism, good health, and aesthetic appreciation – tend to accrue disproportionately to those who, by virtue of their employment, have a steady source of income.

Until recently, however, most of the effort expended on developing data systems has not included data on income and workforce participation. For example, only about half of the documented state student unit-record (SUR) databases in postsecondary education have been linked to unemployment insurance (UI) wage record files, and many of these instances were initial, one-time efforts. While there is still much work to be done in linking K-12 and postsecondary records, states also should be planning now for how to incorporate workforce data into their longitudinal data systems. Indeed, the federal government has made this a basic expectation for states receiving ARRA funds.

Although technological advances now make it more feasible to link data than before, a host of other obstacles must be overcome to ensure successful and full deployment of an HCDDS.

- Data alone, even if they are "accurate," are meaningless without a context to turn them into useful information. Policy and practice must guide these efforts. A purely technical solution to assembling the data linkages may not provide the necessary information. It is imperative to have a clear understanding of the specific policy questions that an HCDDS could address. Those policy questions need to drive the design of an HCDDS from the outset and be constantly revisited as it is developed.
- Good information can be threatening to those who perform below average. By definition, half of the cases will share this fate.
- Data system development is determined, in part, by a given state's historical experience (what is counted is counted in part because it was counted last year). Systemic change is difficult, particularly in times of scarce resources.

Successfully surmounting such challenges requires an intentional process of building buy-in and giving comfort to the various entities who own the data that are to be included. Doing so demands that the individual and collective needs of each participating sector be acknowledged and addressed.

Sector Motivations

The development and subsequent use of an HCDDS can be an expensive, complex political and organizational endeavor. It is therefore important from the outset to identify and clearly understand the specific ways each of the three principal sectors will benefit from their involvement in building a HCDDS. Understanding these motivations also can help ensure that the database is designed to address the proper questions. Generally speaking, the following are the most significant benefits available to each sector.

- The K-12 sector wants information about collegiate performance and job placement so it can improve the effectiveness of curriculum and pedagogy in preparing students to take a next step.
- The postsecondary sector wants information about work placement and earnings to improve the effectiveness of its instruction – especially in programs oriented toward vocational or professional preparation. In parallel, it wants information about prior K-12 achievement to identify areas in which better secondary preparation is needed. Such information can be useful in forming or enhancing partnerships with particular schools and districts, in order to collectively help students succeed prior to high school graduation (e.g., the California State University Early Assessment Initiative).
- The workforce sector wants information about prior training in high school and postsecondary institutions as a foundation for working with both education sectors to address identified skill gaps in the workforce, as well as to identify equity gaps with respect to demographic representativeness by job category. Knowing the education sectors' capacity to respond (i.e., by increasing the flow of graduates with particular skill sets) will also help the state decide whether to invest in education to address skill gaps or establish incentives to induce more workers with needed skills to move into the state. Moreover, linking with the education sectors would provide labor market analysts with a wealth of data that would be useful for examining equity in employment.

The Policy Questions of Interest

The specific data element contents and analytical capacity of a given HCDDS ultimately will be determined by the kinds of questions it is designed to answer. Since many policy questions can be answered without linking data, this section focuses on the kinds of questions that require data drawn from two or more sectors or two or more states.

The figures below provide a conceptual picture of the ways databases can be linked and offer a way to classify different kinds of policy questions. We look at five different questions, ordered in terms of increasing complexity with regard to the combination of data systems needed to obtain answers:

- Questions involving only **one database in a single state**.
- Questions involving **two databases in a single state**.
- Questions involving all **three databases in a single state**.
- Questions involving **several similar databases across multiple states**.
- Questions involving **multiple databases and multiple states**.

Figure 1. Questions Involving One Database in a Single State
(Example: What proportion of students beginning college in Oregon earn a bachelor's degree in six years?)

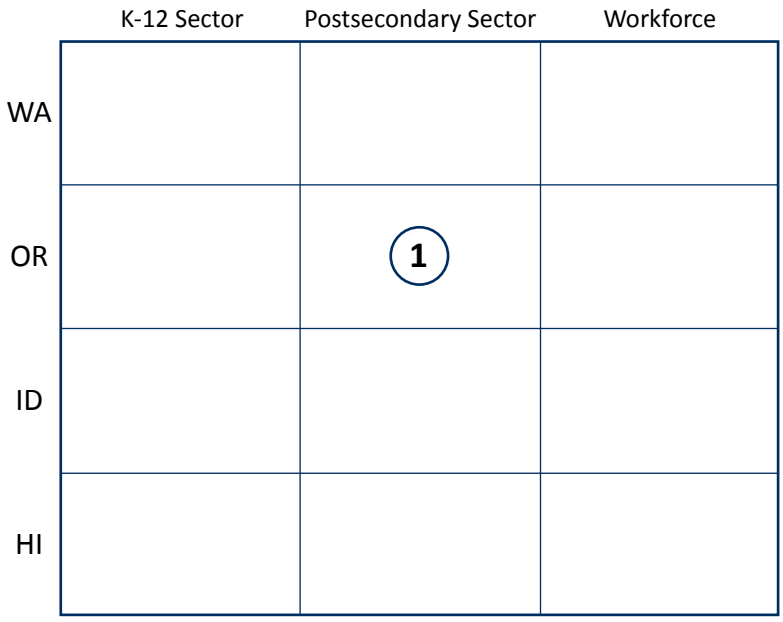


Figure 2. Questions Involving Two Databases in a Single State
(Example: What proportion of students completing high school in Hawaii enroll in college in the state within a year?).

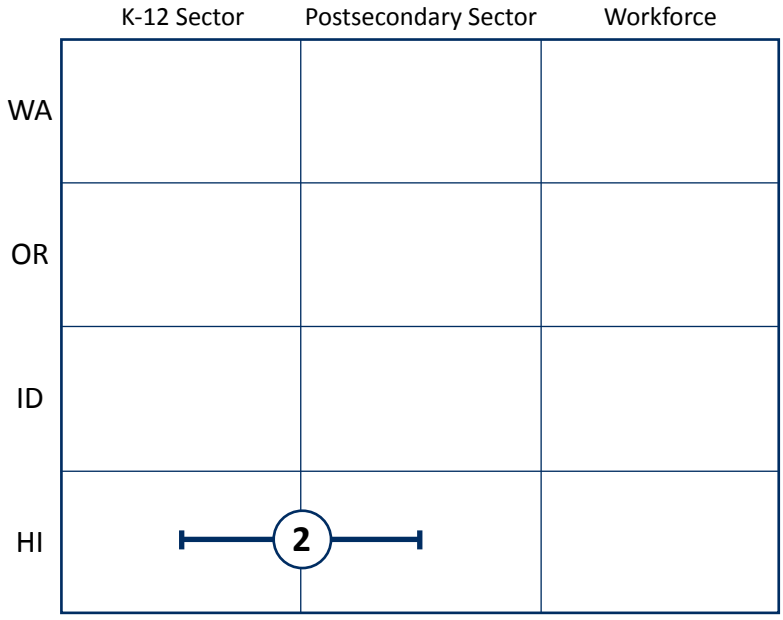


Figure 3. Questions Involving all Three Databases in a Single State
(Example: What proportion of high school graduates in Washington complete college within 10 years and are earning $35,000 or more per year?)

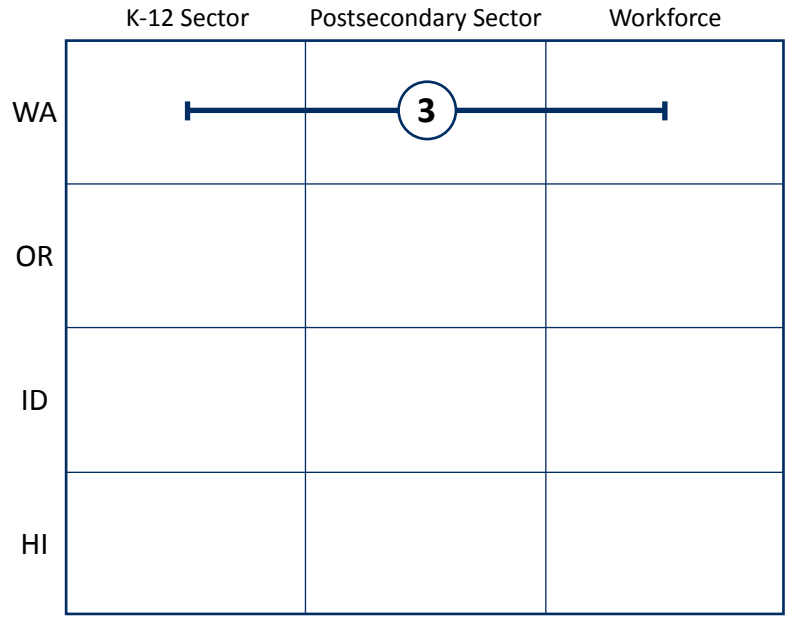


Figure 4. Questions Involving Several Similar Databases across Multiple States
(Example: What proportion of students who were enrolled in college in Washington in a given year are enrolled in Oregon, Idaho, and Hawaii the next year?)

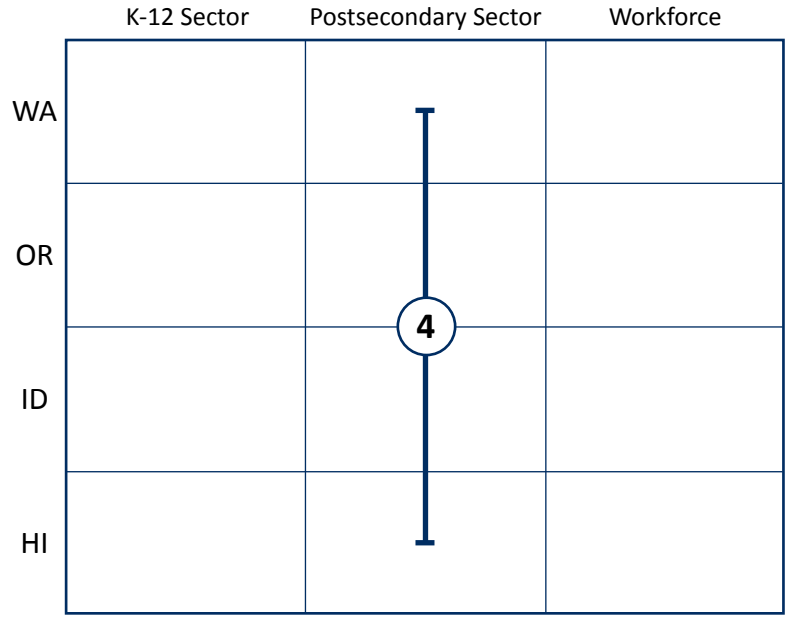
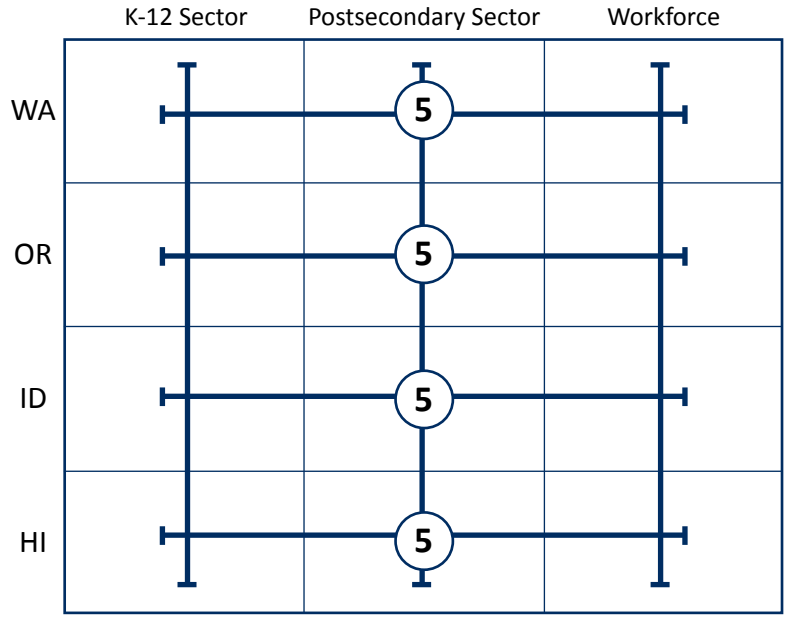


Figure 5. Questions Involving Multiple Databases and Multiple States (Example: What proportion of students completing high school in Idaho complete at least an associate degree and are employed in the aeronautics industry in the state or in Washington, Oregon, or Hawaii?)



Data sources currently exist to answer some of these questions. For instance, the first question is already a standard part of the federal IPEDS. One also can research the second question for any state through IPEDS, although it is only possible to obtain aggregate information. Researchers would have to locate other sources to disaggregate that question by race/ethnicity, age, or any other characteristic. Some states that have invested resources in building a data infrastructure can provide answers to the second and third questions. With external funding and assistance from national and regional organizations, other states are currently making headway on building that capacity. Existing databases or recent development efforts, such as at the National Student Clearinghouse, can help provide answers to the fourth question, although these horizontal data-sharing capabilities do not have a long history, especially in K-12. Progress clearly is being made, but much more needs to be done to address the second through fourth questions. An HCDDS adds the capability to address the fifth question.

Within this framework, the detailed policy questions that an HCDDS should be able to answer are of two principal types. Each of these "master questions" can be further disaggregated to yield dozens of derived questions that address different populations or regions.

Master question 1. How are former high school students from participating states performing in postsecondary education in participating states?:

- Within a certain time period?
- By school and institution attended?
- By key demographics (e.g., gender, race/ethnicity)?
- By type of high school curriculum or particular classes taken?
- By level of readiness for college (developmental placement)?
- By field of postsecondary study (Classification of Instructional Program (CIP))?
- By different departure conditions (e.g., no diploma, GED, high school graduate)?
- By different postsecondary enrollment conditions (e.g., receiving financial aid, full-time/part-time study, non-credit participation)?
- By different postsecondary completion outcomes (e.g., graduating, not graduating)?

Master question 2. How are former high school and postsecondary students from participating states performing in the workforce in participating states?

- Within a given time period?
- By school or institution attended?
- By field of postsecondary study (CIP) or type of high school curriculum?
- By industry of employment (Standard Industry Classification (SIC))?
- By key demographics (e.g., gender, race/ethnicity)?
- By region within state?
- By different departure conditions (e.g., graduated/not graduated, number of postsecondary credits earned by time of departure)?

Many of these sub-questions can be combined to yield queries about more complex topics— for example, the income gain (and consequent increase in state tax revenue) experienced by high school versus postsecondary completers employed in various health-related fields; or the effectiveness of particular high school course-taking patterns in preparing underserved students for successful postsecondary study in the sciences. As the framing of the master questions reveals, moreover, they can also be posed and answered within a given state or across a multi-state region.

Data Contents and Database Organization

The data contents of an HCDDS follow directly from the policy questions outlined above. Relatively few data elements will be needed from each sector because only those that add value for collective purposes should be shared or maintained. A longitudinal data system, as envisioned here, could conceivably be used to answer all manner of questions beyond the two principal ones set forth above – including questions that address causality of specific educational interventions – if the system is populated with enough variables.

But the effort to add data elements can be significant and even controversial. States might consider whether adding variables might slow or sidetrack the effort to construct such a system or threaten its effective usage in answering the two principal questions. It is likely that the data needed to incorporate explanatory variables into a research design could be obtained in other ways that are unlikely to greatly distort the analysis.

Once the requisite datasets are linked or merged, each user community will be able to access its own detailed data holdings, in order to disaggregate results further for particular populations or to conduct detailed cause-and-effect studies. For instance, K-12 users will only need to know a few things about postsecondary enrollments – like the institution at which its former students are enrolled, their majors, and information about a few areas of academic performance – because they already have detailed data about student demographics and classes taken in their own databases. Once the postsecondary performance information is linked in, all these additional variables can be harnessed. For example, a study of the effectiveness of a technology-mediated curriculum could rely on an institution's internal data system to supply needed explanatory variables. Incorporating information from the HCDDS might enhance such a study by allowing the researcher to account for the outcomes of students who disappeared from that local database, but it is probably not necessary for the study to include all students in a state in order to obtain useful findings. Other examples showing how an HCDDS need not be all-encompassing to provide value are easy to imagine.

Data elements are of three basic kinds:

- Performance or outcome: describing the behavior or attainment of individuals in each sector.
- Descriptive: distinguishing members of different populations or different experiential or treatment groups (distinguished by features such as race/ethnicity or the need for remediation, for instance), chiefly included to enable disaggregation.
- Key links: data elements with attributes that enable information for the same individual or entity to be merged. Since not all individuals have adequate unique identifiers, demographic variables like date of birth also may be used in combination as key links.

The following represent the most basic set of data elements needed in an HCDDS to address the most important policy questions.

K-12 (one record per school attended per year)

- Student identifier
- Date of birth
- Gender
- Race/ethnicity
- Free and reduced lunch indicator (or other indicator of income)
- High school attended
- County of origin
- State of origin
- Upper-level math course (record each occurrence)
- Upper-level science course (record each occurrence)
- AP Course (record each occurrence)
- State exam score (latest)
- GPA (term and at graduation)
- Graduation/completion flag (together with diploma type)
- Award date

Postsecondary (one record per academic term)

- Student identifier
- Institution attended
- Year
- Term
- Date of birth
- Gender
- Race/ethnicity
- Income/receipt of Pell Grant
- State of origin
- County of origin
- Student class level (e.g., freshman)
- Full-time, part-time indicator
- Remedial course placements, enrollment, and completion, by subject area (record each occurrence)
- Current major
- Level of degree/certificate completed
- Degree field of study

Workforce (one record per quarter from UI wage record)

- Worker identifier
- Employment status
- Wages earned
- Industry in which employed

This list includes data elements included in the UI wage record and those needed to address the vast preponderance of policy questions that are likely to arise. Additional data on key courses taken in high school and in the first year of college might allow researchers in both sectors to better focus their instructional efforts but would seriously

complicate the design of the HCDDS by changing the primary unit of analysis from person to course enrollment. In the short term, such detail would probably not be worth the effort, unless it could be obtained quickly and cheaply.

As states move forward in creating linked data systems, it is vital for them to attend to the issue of standardization of data elements. To borrow a frequently used analogy, our nation's railways could not operate efficiently if the tracks did not share uniform characteristics. While nothing as dramatic as a train wreck will happen if data elements purporting to measure the same thing do not share a common definition, a lack of standardization among those elements is a major barrier to interoperability and cogent analysis.

As it stands, states have defined similar concepts in slightly different ways, and that will surely bog down – if not completely derail – work to link data systems. It is vital that core data elements be standardized in the early stages of longitudinal database design. Careful attention to the standardization efforts taking place across the country is crucial, lest we find ourselves with 50 (or more) incompatible data systems and no straightforward way to analyze data from more than one state. Data standards regarding each variable's "technical" construction (i.e., whether it is a string or a numerical variable, how it is coded) and "business" construction (i.e., how race/ethnicity is defined) must be developed.

Although this discussion treats an HCDDS as though it were a single database, there are many different ways such a resource might be constituted and organized. They include:

- A single merged database that is a physically maintained entity separate from any of the individual sector databases and is intended to replace them. This database would include all of the variable content of the original sector-level databases in a fully relational organization. Such an arrangement has the advantage that any relational question can be posed and answered, but it has the major drawbacks of costing a lot (both in direct costs and in the opportunity costs entailed in scrapping existing capacity) and being difficult to document and maintain. The Florida Education and Training Placement Information Program database is the closest extant state data resource to this configuration.

- A far smaller merged database that contains a "common core" of data and that is a physically maintained entity separate from each of the individual sector databases. The data content of the common core would correspond to the list presented above, with data elements extracted from the parent databases according to a defined schedule. Users in each sector could access data elements drawn from the common core via a unique identifier and could link them into analytical files containing outcome variables and a range of other data on the same individuals, drawn from their own data

records. This approach is cheaper and easier to maintain than the above alternative, but it may not be able to address all potential policy questions. This approach was used in the NCHEMS multi-state data exchange demonstration project in postsecondary education. On a more limited basis, it is also how the many examples of bilateral matches (e.g., K-12 to postsecondary or postsecondary to workforce) have been accomplished in various states.

- Clearly established "gateways" or "paths," by means of which authorized users in each sector can access a limited set of data elements (probably the common core, above) directly from one another's databases. This would avoid the need to create a separate, mutually accessible database, as well as the awkwardness of one agency giving up direct control of some of its own data by moving records to a third-party database. There might be FERPA (Family Educational Rights and Privacy Act) advantages to pursuing this route. Operating in this way, however, would create linkage pathways that might be difficult to maintain and keep secure. No examples of this architecture currently exist.

- Special databases created for analytical purposes on an ad hoc or bilateral basis with data drawn from parent databases in the various sectors, as needed. This option has few advantages because capacity has to be reestablished every time a new policy question is posed. A number of states have experimented with cross-sector data linkages, however, so quite a few examples of this approach exist.

On balance, the second approach is probably the most feasible, given available funds and talent. It is also the alternative for which there is the largest body of extant experience.

Technical concerns relating specifically to how such a data system might be constructed should be considered last. These issues should be least influential in driving the design and governance of the data system. Nonetheless, technical issues are important once larger questions of motivation and use are settled. In particular, the means by which states will match individual students in the face of different procedures, laws, and regulations governing the use of common identifiers will constitute a technical challenge. Perhaps the biggest challenge is matching K-12 students who are not found in one of the states' postsecondary databases with available employment records if the K-12 system does not collect student Social Security numbers or is prohibited from doing so. Given the sensitivity of SSNs and the fact that even they cannot match all individual student records "perfectly," it is probably wise for states to adopt a broader approach to "identity matching." Such an approach would link records using a larger group of variables corresponding to student characteristics, including but not limited to the SSN (when available) or statewide student identifier.

Governance Arrangements

Establishing effective governance structures and procedures for any linked data system is a topic that demands careful attention. Making a data match, though a technical challenge, requires only a small fraction of the time and effort needed to establish an HCDDS. Questions related to who owns the data, who has the right to use it, how data quality is managed and assured, and how to merge databases despite inconsistent definitions for otherwise similar variables can be the subject of endless debate. Analytical interpretations of obtained results, once the data system is in place and is being used, can be another source of tension. If governance issues are not addressed at the outset, a data system is unlikely to be developed at all. Therefore carefully working through these governance issues early on and continually revisiting them will help ensure that the data system remains vital and useful.

Governance issues within any sector in any state are difficult enough. Establishing these arrangements properly becomes even more crucial to the success of a multi-sector, multi-state data exchange. Under these circumstances all the parties are present voluntarily and must be continually reminded of the benefits that they derive from involvement. There is no governor or legislature to mandate cooperation. In fact, because performance comparisons across states are inevitable once such a data resource is established, there may be built-in disincentives to collaboration in the first place. So the results produced by such a system must be especially compelling and its governance arrangements especially sound.

Within states, experience has shown that there are several workable solutions to establishing a governance structure. Among them are the following:

- Establish a set of inter-agency memoranda of understanding (MOUs), allowing each party access to a limited set of data elements on a periodic basis. This is the way the majority of cross-agency data-sharing arrangements have been organized within states to date.

- Establish a "lead" agency to take responsibility for making the match and maintaining the resulting data. Current FERPA regulations suggest that this agency be an education agency whenever workforce data is accessed. Texas, among other states, has adopted this approach, using the Texas Higher Education Coordinating Board as the lead agency, which physically links UI wage-record data to educational records in its own facilities and under its direct control.

- Establish a new agency or authority by state law to assume these responsibilities, as was done in Florida. It is frequently a good idea to establish explicit authority to create an HCDDS through state legislation, regardless of how it is organized.

Fortunately, there are also models for making governance work across state lines, at least for a short-term analytical project. NCHEMS put

together such an arrangement in an effort to determine how students moved among postsecondary institutions in four states. The central feature of the governance arrangement that made the data exchange possible involved NCHEMS crafting multiple bilateral memoranda of understanding with the states involved. (See the box on page 16 for a link to the report from that study, which includes an example of the MOU negotiated to share the necessary data.) The four states engaged in this pilot effort first proposed a general approach to the governance of their prototype exchange by blocking out associated roles and responsibilities. They then conducted a series of extensive constituency consultations, involving institutions to ensure that the motives of the undertaking were clear and the limits on what could and could not be done were established. Following these consultations, bilateral agreements between NCHEMS and each participating state were crafted, each tailored to the specific, and different, legal requirements of the state. This general approach should be followed in crafting future multi-sector, multi-state data-sharing arrangements.

A permanent multi-sector, multi-state HCDDS will require a robust governing council to establish appropriate policies, related to issues such as who can gain access to the data under what circumstances, how long data should be archived, how privacy should be protected, and limits on appropriate use. The council should consist of at least one representative from each of the involved sectors within each state (e.g., K-12 education, postsecondary education, and workforce development/ labor). Each state delegation might identify one among their number to serve as a lead. These leads would form an executive committee within the governance council, which would meet more regularly. A parallel technical body might also be established to set and change data element definitions and provide guidance on how data can be most effectively used.

Once the development phase of the data system is complete and the funding associated with development runs out, the original governance arrangements might be too expensive to continue. Under these circumstances, the executive committee might suffice as a governance structure. This approach presumes that the data exchange has evolved in a climate of trust and that these arrangements have in fact generated mutual benefits for each of the sectors. It realistically balances the need for each state to have a voice in the oversight of the system with a reasonable cost for maintaining the structure.

The governance structure will also manifest itself in the architecture of the eventual system, as discussed above. Several models exist. They include:

1. Warehousing all relevant data with a third-party organization. This has the advantage of simplicity and means that the participating state agencies will not need to employ staff, but it would entail

permanent costs associated with the operations of a third-party organization.

2. Outsourcing the match (the actual task of linking the student unit records together) to a state with a proven capacity to do so. Advantages and disadvantages of this approach are the same as those above, with the additional disadvantage that the states, other than the state to which the match was outsourced, give up direct control of their data to a "foreign" state government entity.

3. Broker-based matching, as was done in the NCHEMS four-state demonstration project, in which a trusted third party is tasked with identifying and managing the work of an outside technical resource with the capacity to make the matches. In this case, the broker provides regular reporting and possibly meets specific ad hoc analytical needs. This has the same advantages and disadvantages as the first alternative. It may also include slightly higher costs, since there still would be the need for an organization to perform the actual matchmaking in addition to the broker role. But such an arrangement has the advantage of not conflating the roles of the technicians performing the match with the policy experts who serve as the brokers and principal analysts.

4. Multilateral or multiple bilateral MOUs among the states. At the outset, this entails the least cost and is the most flexible, but it requires attention to maintaining a plethora of different agreements across agencies and states.

Security and Privacy Considerations

Access to and use of educational data for both K-12 and postsecondary agencies is regulated by FERPA and may be additionally governed by state law. Parties attempting to set up an HCDDS should carefully review any relevant legislation that might affect their approach. Privacy law is another good reason why authority to proceed should be sought through specific state

Additional Resources

- NCHEMS, "Tracking Postsecondary Students Across State Lines: Results of a Pilot Multi-State Data Exchange Initiative" (Boulder, CO: NCHEMS, 2008): http://www.nchems.org/c2sp/documents/ResultsofMulti-StateDataExchange.pdf)
- Model MOUs are available in the appendix to the NCHEMS study
- Data Quality Campaign's resources on FERPA compliance: http://www.dataqualitycampaign.org/resources/topics/13
- Statewide Longitudinal Data Systems Grant Program information from the Institute of Education Sciences: http://nces.ed.gov/programs/slds/
- Mills, Jack, "State Data Systems and Privacy Concerns: Strategies for Balancing Public Interests" (Boston, MA: Jobs for the Future, 2005): http://www.jff.org/sites/default/files/StateDataSystems.pdf

legislation directed toward the participating agencies. Respected legal opinion holds that FERPA permits all of the steps needed to created, load, and use an HCDDS, provided that certain guidelines are followed. There is a sound literature on these that is incorporated into this framework by reference, and several resources are provided in the box on this page.

Data security is also an important consideration. Database arrangements must be physically secure, with password-protected access to data confined strictly to designated authorized users within each agency. This means that establishing physically separate databases may be better than linkages and pathways purely from a security perspective. Encryption is another important tool in ensuring data security and should be used whenever data are moved from one location to another electronically. Finally, privacy and security considerations may require the perturbation of results if cell sizes in any analysis fall below five cases.

Conclusion

A multi-state data exchange – what we have chosen to call a human capital development data system – that enables policymakers to look comprehensively at the stock and flow of human capital has become essential for effective policymaking and planning in the globalized knowledge economy. Technology now permits the development of longitudinal systems that follow individual students from elementary school through college or directly into the workforce, but existing systems, organizational boundaries, and governance issues all present formidable barriers. This paper argues that the development of longitudinal data systems should be guided by two basic questions only, while allowing for meaningful disaggregation to examine how policies and practices may be disparately affecting individuals based on race/ethnicity, income, or other characteristics. It may be unwise to seek some form of an "ideal" data warehouse with all conceivable data elements, at the risk of disrupting momentum toward development of a robust longitudinal data system.

An HCDDS can also enhance the analytical power of more traditional sample-based research designs that seek to establish causal relationships. The inclusion of multiple states in a data exchange surely presents additional challenges, but the need to incorporate information related to individuals' mobility across state lines is great. There appears to be both interest in and promising models by which to design a governance structure that meets this need and can be put in place on a permanent basis. To work, data element standardization and interoperability is a key ingredient that demands attention early in the developmental process.

Acknowledgments

The authors would like to thank the participants at the June 29, 2009, meeting in Olympia, WA, for their contributions to the discussion about how to develop a framework for a multi-sector, multi-state data exchange. Special thanks also goes to Jeff Stanley and Hans L'Orange at the State Higher Education Executive Officers for their feedback on prior drafts of this document and to Annie Finnigan and Candy Allen at WICHE for their editing and graphic design expertise. Our gratitude also goes to the Bill and Melinda Gates Foundation, which provided funding for this paper.

Endnotes

[i] Some of the data elements in this list will be easier to obtain than others. Particularly nettlesome are the measures related to income or socioeconomic status, and they deserve special mention here. At the K-12 level, schools are required to provide information on students receiving Free and Reduced Price Lunch (FRPL), while the most commonly used indicator at the postsecondary level is receipt of a Pell Grant. Unfortunately, research suggests high school students are less likely than students in lower grants to be counted as FRPL recipients. Likewise, since students must complete the Free Application for Federal Student Aid (FAFSA) to be eligible for a Pell Grant, it is an imperfect indicator of income. In both cases the dichotomous nature of these data elements compounds the problem, since the failure to correctly identify a low-income student means he or she is misidentified as not low-income. FAFSAs do provide more nuanced income information, but analysts or data collectors have few tools for accurately imputing income for those students who do not complete FAFSAs. Because access to financial resources is a key to enrolling and succeeding in college, we include measures of income in this core list of data elements.

[ii] "Current major" differs from "degree field of study" here only in that the former field would be populated each time a student change his or her major, whereas the latter field would only be populated when the degree or award is conferred. At that time the two fields would most likely match.

For further information contact:
Brian T. Prescott, Ph.D.
Director of Policy Research
Western Interstate Commission for Higher Education
3035 Center Green Drive
Boulder, CO 80301-2204
P : 303.541.0255
or bprescott@wiche.edu