Data Exchange Data Pathways

As the data exchange project works toward building a sustainable governance model and a set of Memoranda of Agreement among and between states for an initial exchange of data, there is a need to lay out more specifically the different possibilities for how data might flow among education and labor market data sources for linking and analysis. The following exhibits are an attempt to lay out the possible pathways for data to flow within a data exchange among states’ education and workforce sectors, a matching vendor, and national/federal databases with great potential for expanding the reach of the information that can be shared. All exhibits assume that the data exchange team will collectively develop a definition for membership in the cohort or cohorts for which data will be exchanged, linked, and analyzed. This discussion is a reflection of what we have learned to date through our discussions with participating states, the National Student Clearinghouse (NSC), the Census Bureau, and the Department of Labor; a review of documentation, including legal agreements among states and within states for data sharing across sectors; and past experience. It also assumes that the recently issued FERPA guidelines will adopted essentially as proposed in April 2011.

These exhibits are not meant to be exhaustive of all combinations for linking these data; rather, they are intended to be illustrative of approaches that differ in ways critical to what the eventual data exchange’s capabilities are. After a brief description of the process illustrated by each exhibit, a discussion briefly describes the tradeoffs that accompany each model, in the form of a brief list of advantages and disadvantages.

Generally, each of the exhibits depicts a scenario in which each participating state is expected to craft the initial cohort in identical ways (or not craft one, as is the case with the first exhibit). But in reality, individual states have varying capabilities for linking data within their own state, not to mention other possible reasons for why they may prefer a different level of pre-matching engagement than their neighbors. So nothing in this attempt to describe pathways for data exchange should limit the latitude for states to determine what form their involvement should take, so long as the state can assemble a cohort(s) adhering to the collective definition.

In addition to WICHE and the participating states, the data exchange’s operation requires a capable additional partner who can do the technical work of linking the data from these disparate sources together. The entity fulfilling this role would need to be able to provide the necessary level of programming expertise to manage and match the large volume of data anticipated to emerge in this project and to perform analytical work on the resulting dataset, while also having the capacity to both store and appropriately secure the data. In effect, the data exchange could rely on any vendor with the requisite capabilities to perform matching, including for- or non-profit entities or even, as some have suggested, one of the participating states willing and able to take on this responsibility. The data exchange’s participating states need to have comfort that the vendor assigned to perform the matches is reliable, secure, fully qualified to do the work, and cognizant of the sensitivity of the data entrusted to its care.
One potential choice of vendor to serve in the matching role is the NSC itself. With a long history of storing and linking longitudinal student records, the NSC meets the required capabilities well and is already involved in the kind of work the data exchange is trying to establish. Moreover, in its own collections of secondary and postsecondary students, it represents a resource for capturing efficiently many of the pieces of information that the data exchange will need to fulfill its potential.

But it is essential to understand that we envision these two roles – that of data source and that of matching vendor – as very distinct ones. If NSC fulfills the role of matching vendor, these roles can easily become conflated. But the NSC has been an active participant in our discussions to date, so it is worth carefully clarify these role and the exhibits below attempt to distinguish them, except where otherwise noted.

The NSC’s first and more broadly understood role is as a potential data source through matching state-provided information with its existing data collection that it acquires from individual postsecondary institutions (and, to a lesser extent, public school districts). With respect to our multi-state data exchange, this role can be further unpacked according to whether the NSC’s main collection is serving as the primary data source or as a supplement to a state’s own data. In the latter, the NSC’s data are only used to fill in information about student enrollments and degrees awarded that the state does not already have, and typically concerns activity occurring outside the state’s own higher education institutions. States that have contracts with the NSC can obtain these data for students who they can first identify (such as for individuals who transfer out of an in-state public institution) and use them for their own analytical purposes. The NSC can potentially serve as a data source regardless of whether or not it fills the matchmaking role.

As a potential matching vendor, based on prior discussions with the NSC, its plans to serve as a potential matching vendor include the creation of a separate “partition” of its data storage system to contain and manipulate exchange data. To the extent that the exchange is able to obtain data from NSC’s main collection and use them fully, those data could be brought over from NSC’s main collection. Otherwise, the exchange dataset will be “walled off” from the rest of the NSC collections. This would allow the exchange to take advantage of the NSC’s technical know-how, its extensive storage capacity, and its data security apparatus without having to mingle the exchange’s data with the NSC’s main collection.

General Process and Key Questions

In this discussion it is important to note that the project is committed to exploring the extent to which a regional data exchange approach can 1) make possible a multi-state view of the creation, stock, and flow of human capital and 2) provide each participating state with enhanced data about the educational and labor market experiences of their own residents obtained from neighboring states. Accomplishing the former is possible simply by linking the data across state lines and producing relevant aggregate analysis and reports. The latter may require the exchange to be able to deliver a data file (either identified or de-identified) with enhanced individual-level data back to the state from which the original records came.

Speaking in general terms, there are six steps involved in accomplishing what we have set out to do with this multi-state data exchange. These are not necessarily sequential steps, and there may be variations
in the order and manner for different states, and the process may even leave out one or more steps or require additional ones depending on state capabilities and other considerations. But for a data exchange to function long-term, states will need a way to:

1. Link public education records (K-12 and postsecondary).
2. Enhance education records with NSC data to capture activity in out-of-state and nonpublic institutions.
3. Link to workforce records (generally Unemployment Insurance (UI) wage record files.
4. Merge records across states.
5. Analyze and report on the resulting dataset.
6. Return a data file containing the original records enhanced through this process to the originating states.

Other steps such as linking student-level data with other state sources, such as the Department of Motor Vehicles or health and human services or welfare agencies, may also be helpful (especially to obtain a social security number required for matching the UI wage records) or desirable.

In building a data exchange and the MOA documents that animate it, there are several key questions that need answers, including:

1. Whether and at what point to integrate data from NSC’s main collection?
2. Is the matching vendor one of the states, the NSC, or another third party entity? This question is discussed above.
3. Do states link data before sending it to the matching vendor? If so, do they link just their educational data or both educational and labor data?
4. What data sources should provide the labor data? Here the question is asking whether the exchange should get those data from each state individually or from existing federal data collections, namely Census LED and WRIS/WRIS2.
5. How can we accommodate different state needs/desires and different capacities to contribute merged data?

By putting forth a discussion of the options, this document intends to help lead to answers to the above questions.

**Notes about the Labor Data Match**

Linking in workforce data is essential to the success of this data exchange. While the underlying data for performing the match comes from the Unemployment Insurance (UI) wage records, there are three major sources for gaining access to and linking in those records. The first is simply to query each state’s own UI files. As explained in the exhibits below, this approach may allow for the greatest flexibility in making the full dataset subsequently available for research or for supplying each participating state with enhanced workforce data. It may also best approximate a true “exchange” of data across state lines. Under this scenario, participating states (and their agencies) have the best chance of being able to fill in gaps in their existing data for students whose choices have led them across borders, while still remaining
compliant with FERPA and other applicable privacy laws and regulations, if they are exchanging the information directly with fewer intermediaries.

But there are also two federal data sources of UI files. First is the Local Employment Dynamics program at the U.S. Census Bureau (Census LED). It compiles quarterly UI wage record data from all 50 states, D.C., and Puerto Rico into a single data resource, and then supplements that information with data from its own and other federal sources. The resulting file includes not only the original UI wage records, but also information about an individual’s demographic characteristics, including race/ethnicity, age, etc. Census LED has strict confidentiality requirements, so matching with this resource means the capability of querying the data for use by participating states is limited; once the data are matched, only aggregate reporting is possible. However, it may be possible for researchers working under approved research projects to access the resulting matched data at one of the Census Bureau’s regional data centers located in Atlanta, Boston, Charlotte, Chicago, Dallas, Denver, Detroit, Kansas City, Los Angeles, New York, Philadelphia, and Seattle.

The second federal data source is the Wage Record Interchange System (WRIS), supported by the Education and Training Administration within the U.S. Department of Labor. WRIS is a cooperative agreement among all 50 states to exchange UI wage records among one another. Essentially, it provides a crosswalk table of Social Security Numbers, which a state can query with their own valid SSNs to discover if those individuals are found in other states. It is useable only to evaluate workforce training programs under the aegis of the Department of Labor. WRIS2 is second system that extends WRIS to include other programs that prepare individuals for the labor market, including secondary and postsecondary education programs. WRIS2 is still under development, and only a handful of states are likely to be signatories in the early rollout period. In both cases, the data sharing agreements stipulate that the only organization permitted to access the system are known as Performance Accountability and Customer Information Agency (PACIA), which are state agencies designated by the governor. PACIAs would be required to perform all data analyses using WRIS data and provide only summary information back to the requesting entity. Under neither WRIS system does it appear at the moment to be an option for enhancing states’ original data files for their subsequent use and analysis.

Other Notes

There are several assumptions embedded in the exhibits and discussion below. First, it is understood that the research agenda and the research questions to be pursued are under the control of the data exchange’s governance body in whatever form it may take. The governance body also will specify the cohort(s) definition to be used for extracting the original data files each state will submit to the matching process. Any standard reports would also be approved by the governance body.

Second, it may be that one or more of the data suppliers to the exchange will be unable or unwilling to provide data at an equivalently detailed level. For example, the NSC may be resistant to identifying the institution where a student is enrolled. This discussion assumes, however, that there exist intermediary levels of specificity in data that can satisfy all partners, and that it is not necessary for data to be exchanged on an all-or-nothing basis. Rather than leaving fields blank, and the data exchange process no
richer for it, it may be that the partner could recode or impute information in that field in a way that still enhances the capacity of the exchange. Using the previous example, the NSC could perhaps recode information in the institution field so that the institution is not individually identified, but information about the student’s enrollment in a particular sector or state does become available to the exchange where otherwise it would not have been.

Finally, in order to further shield individual records, it may be necessary for the core exchange dataset to be stripped of individually identifiable information and an exchange ID created for the purposes of making the full dataset available for analysis. A crosswalk table may be kept in order to recreate the matches so that any questions or concerns about the quality of the matches can be addressed, but access to that table would be more secure and restricted.

It goes without saying that all arrangements will be subject to review and approval by the requisite legal authorities within each state. The data exchange will remain in full compliance with federal and state privacy laws.

Exhibit 1.

Exhibit 1 illustrates the simplest pathway for data exchange. It would take full advantage of existing national/federal data stores encompassing a majority of individuals enrolled in secondary and postsecondary education, and employed in the labor market to complete a research matrix for participating states. The secondary and postsecondary data would come from the National Student Clearinghouse, which contains __% of secondary enrollments and about 92% of postsecondary enrollments. Labor market data would come from either the Wage Record Interchange System (WRIS), WRIS2 (an extension of WRIS that is still in development with some education and training data but a
minority of participating states), or Census Local Employment Dynamics (Census LED), all of which have Unemployment Insurance (UI) wage record files. In this model the NSC would play a driving role simultaneously as the supplier of information and the matching vendor by constructing the relevant cohort(s) out of its own records and linking them to the labor data.

**Advantages**
- Straightforward, simple
- Takes fullest advantage of existing data stores
- Incorporation of non-participating states data relatively easy
- Potentially limits the degree to which state agencies need to commit resources
- Has information on private postsecondary enrollments

**Disadvantages**
- NSC’s collection for secondary enrollments and postsecondary degrees is currently much more limited than its collection of postsecondary enrollments
- Limited scope of data elements available
- Generates no buy-in from the states
- No opportunity for states to obtain enhanced unit record data files, either for postsecondary enrollments or labor data, since both NSC and the UI suppliers will insist on providing only aggregate data
- Probably not possible to link labor market information in via this approach, due to lack of SSNs except those within the NSC’s main collection

Exhibit 2.

In Exhibit 2, states compile separate data files for their K-12 and postsecondary education sectors, based on the exchange’s agreed-upon cohort definition, and send those to the matching vendor to merge them all together into a “core” exchange dataset. Once the core dataset exists, the matching vendor will
send the core dataset to one of two places for matching with labor market data. Either the matching vendor will send the dataset to Census LED, which will perform the match and produce the aggregate report.¹ Or the matching vendor will send the core dataset to a participating state’s PACIA for matching files to the WRIS or WRIS2 systems. The PACIA will subsequently produce the aggregate reporting. The aggregate report will be the only product of the exchange.

Advantages
- Involves the states in cohort development, potentially helpful in a few ways: they get to make judgment calls about the students they most care about within the broader definition
- Takes full advantage of existing workforce data sources that cover the nation, not just participating states
- Census LED querying potentially possible via regional centers
- Captures postsecondary records that NSC may not have

Disadvantages
- Aggregate reports
- Census LED access via regional centers (and thus the nature of the research) is subject to as-yet unknown Census approval process
- No enhanced data files back to originating states
- Match with workforce data cannot be replicated (e.g., it’s a black box)
- Requires navigating vast federal bureaucracies

¹ NSC has discussed the possibility of becoming an “agent” for Census LED and, if that is possible, NSC may be able to query Census LED for the UI records of a select group of individuals (those who are in the core dataset) and then match them internally before producing the aggregate report.

Exhibit 3.
Exhibit 3 differs from the preceding model only in that the states will assemble their cohorts, based upon the exchange’s agreed-upon cohort(s) definition, and merge their within-state educational records prior to sending it off to the matching vendor for merging into the core exchange dataset. Included in the states’ data files will be all available SSNs. In the process of creating the core dataset, the matching vendor will also further enhance it with records from the NSC’s main collection. Subsequently, the matching vendor will send the core dataset to either of the two federally managed sources of UI wage records, through which the labor market data will be appended. The result is an aggregate report, which will be the only product of the exchange.

**Advantages**
- Involves the states in cohort development, potentially helpful in a few ways: they get to make judgment calls about the students they most care about within the broader definition.
- States matching data themselves means only one file goes to NSC
- Takes full advantage of existing workforce data sources that cover the nation, not just participating states
- Census LED querying potentially possible via regional centers
- Captures postsecondary records that NSC may not have

**Disadvantages**
- Aggregate reports
- Census LED access via regional centers (and thus the nature of the research) is subject to as-yet unknown Census approval process
- No enhanced data files back to originating states
- Match with workforce data cannot be replicated (e.g., it’s a black box)
- Requires navigating vast federal bureaucracies

Exhibit 4.
The model illustrated by Exhibit 4 begins with each state applying the exchange’s agreed-upon cohort(s) definition and merging its own within-state educational data before sending the resulting data file to the matching vendor. Included in these data files will be all available SSNs. Next, the matching vendor will match the educational data across all participating states to create the exchange’s core dataset and enhance that dataset with information from its own main collection. In the next step, the matching vendor will send a list of all SSNs to each participating state’s labor market information agency. This agency will append the UI wage record(s) corresponding to each SSN for which a UI wage record(s) can be found and return the resulting file to the matching vendor. Once in receipt of the labor data for all participating states, the matching vendor will merge those files back into the core exchange dataset and use it to generate the aggregate report. The matching vendor will also create subsets of the core exchange dataset that it will provide to each participating state. These enhanced datasets will contain information only for those students the state originally submitted. It may be necessary for the matching vendor to recode individual fields, remove records, or strip out all identifying information in each state’s dataset to prevent identifying individual students.

**Advantages**
- Involves the states in cohort development, potentially helpful in a few ways: they get to make judgment calls about the students they most care about within the broader definition.
- States matching data themselves means only one file of educational data goes to NSC
- Allows for research on the full core exchange dataset (including workforce information), once identifying information is stripped out and replaced by an exchange ID
- Greater flexibility for generating enhanced data files to go back to originating states
- Captures postsecondary records that NSC may not have

**Disadvantages**
- Less efficient in that it requires two separate sets of queries of state resources
- Only allows for labor linkages for individuals working in participating states
- Must persuade additional states to join the exchange or to otherwise make available their workforce records, in order to capture more states’ data
The difference between Exhibit 4 and Exhibit 5 is that, in the model depicted in Exhibit 5, each state merges its educational and labor data before submitting its data file to the NSC for merging. After merging these submitted data files, no subsequent effort is made to re-query the participating state’s UI wage record database.

**Advantages**
- Engages all three principal sectors in each state to submit a single data file for the cohort(s) they collectively assemble
- Captures postsecondary records that NSC may not have

**Disadvantages**
- Prevents matching workforce data across state lines since the only workforce information to be included in original datasets will be for those students who either originated in the same state’s K-12 or in postsecondary education sectors
- May not deliver much more workforce information back to the states than what they originally started with, since the workforce match occurs prior to compiling the core exchange dataset.
- Must persuade additional states to join the exchange in order to capture more states’ workforce data
Exhibit 6.

Exhibits 6 and 7 most clearly illustrate the distinction between the two roles the data exchange’s discussions to date have envisioned for the NSC to fulfill. As illustrated in Exhibit 6, after assembling its cohort(s) according to the exchange’s agreed-upon definition, each state uses its existing contract with the NSC to obtain the enrollment and degree records in non-public and out-of-state institutions for individuals in its cohort file. In doing so, the NSC is serving only in its role as a data source, which requires it to behave no differently than it currently does when responding to similar inquiries from states or school districts with which it has contracts. After merging that file, each state will have a more comprehensive set of information about its residents’ secondary and postsecondary enrollment histories. This data file will also include all available SSNs (though it seems unlikely that, as in all other scenarios, that the file will include any SSNs that could only be supplied by NSC). Next, the states will submit their data files to the matching vendor, potentially the NSC now serving in its second capacity, which will next merge the data files it receives as any other vendor assuming a like role would do, without regard to whether the data in the files originated in the submitting state’s public institutions, the state’s own student unit-record system, or the NSC main collection. Once it has assembled this core exchange dataset, the matching vendor will send the core dataset to either of the two federally managed sources of UI wage records, as depicted in Exhibits 2 and 3, through which the labor market data will be appended. The result is an aggregate report, which will be the only product of the exchange.

**Advantages**
- Separates the distinct roles the NSC is attempting to play
- Cross-state educational matching may mean that NSC can match using all fields supplied to it by the

**Disadvantages**
- Less efficient to query the NSC database separately apart from its activities in fulfillment of its roles as the matching vendor
- Aggregate reports
Exhibit 7 illustrates a model identical to Exhibit 6, up to the point at which the states have submitted their data files to the matching vendor, which has merged them into an initial core exchange dataset. At that point, the process for enhancing that dataset with labor market information follows the path that was outlined in Exhibit 4, with the matching vendor sending SSNs to each state’s UI collection, merging the resulting enhanced data into the core dataset, and producing the aggregate report as well as the enhanced sub-datasets for the individual originating states.

Advantages
- Separates the distinct roles the NSC is attempting to play
- Cross-state educational matching may mean that NSC can match using all fields supplied to it by the

Disadvantages
- Less efficient to query the NSC database separately apart from its activities in fulfillment of its roles as the matching vendor
- Less efficient in that it requires two separate
originating state, possibly including the SSN, even if the state obtained the SSN by way of its initial match with the NSC main collection

- Engages each state in cohort(s) definition, with enhanced out-of-state and private secondary or postsecondary enrollment information
- Allows for research on the full core exchange dataset (including workforce information), once identifying information is stripped out and replaced by an exchange ID
- Greater flexibility for generating enhanced data files to go back to originating states

sets of queries of state resources
- Only allows for labor linkages for individuals working in participating states
- Must persuade additional states to join the exchange or to otherwise make available their workforce records, in order to capture more states’ data