POTENTIAL FUTURE ARCHITECTURES

At the January 2012 meeting in Boise, Idaho, several alternative “architectures” for operating the data exchange were proposed and discussed by participants. These alternatives were raised in the context of a discussion about sustainability of the exchange beyond the end of the grant. We are moving forward with the process as outlined in the agreements we have signed already for the initial exchange of data. Nonetheless, informed by the work we have put in to come to those agreements, we should remain open to other possible architectures and how each might influence how sustainable the exchange may be. With that in mind, the working group suggested that WICHE give some thought to four broad design concepts in particular.

This draft document provides a description of each alternative accompanied by a list, not intended to be exhaustive, of perceived pros and cons. A foundational principle of the data exchange project as envisioned under WICHE’s leadership is that it be used to address policy-relevant research questions identified in advance. Since the expression of research questions as a first step leads directly to the specification of a cohort to be tracked, this principle implies that architectures be cohort-based. Four alternatives are presented here are the cohort-based sequential merge, a fully developed relational database, potentially “plugging into” the core collection efforts of the National Student Clearinghouse, and a WRIS “doppelganger” architecture. WRIS refers to the Wage Record Interchange System, an existing multistate effort to swap individual-level labor market information.

1. Cohort-based sequential merge

This architecture is the approach we are taking to accomplish the initial data exchange via the MOAs making (or have made) their way through state approval processes. It is illustrated in the figure below. The process begins with the extraction of two cohorts from the K-12 and postsecondary education data systems and proceeds in a sequential process as outlined in the MOAs, with each major step identified numerically in order (e.g., step one is to extract the two cohort files and transmit to the matchmaker; step two is to provide merged cohort information to the postsecondary data systems, which will extract matching postsecondary enrollment and awards records; and so on)
Pros:

- Feasible and clearly specified scope: states will only be asked to provide information on the students originally extracted into cohorts, with no need to add new individuals into the dataset at any subsequent point. (although additional cohorts could be extracted at states’ collective discretion).\(^2\)

- Participating states collectively specify which data elements they would like to exchange, rather than be restricted to what a third party already collects or be subject to what other parties (e.g., institutions) are willing to allow to be collected and used.

- Participating states collectively define cohort(s), as well as individually using local knowledge and judgment to determine which students are included in a particular cohort based on that definition.

- Necessary data for students captured as they emerge from high school without neglecting those who begin college at later stages in life.

- Unnecessary to change the design to accommodate the specification of another cohort in a linear/chronological path, such as 9\(^{th}\) graders. Also the addition of another state’s comparable data can fit within this design without modification.

- Adaptable design which will allow for additional research questions that rely on a non-chronological analysis, if desired. For instance, states could define a cohort of postsecondary graduates and look both backwards and forwards in time to see what kinds

---

\(^2\) These refer specifically to the National Student Clearinghouse, which limits the disclosure and use of data according to legal agreements it reaches with individual institutions.
of preparation led to their educational success, and how that success played out in labor markets. Or states could potentially define a cohort of incumbent workers and examine both their prior and subsequent educational paths. These alternative research designs would necessitate a reordering of the sequence in how data are gathered, but such adaptations—with appropriate permissions—could possibly build off of the existing, agreed-upon data sharing process.

- Designated entities that receive personally identifiable information in the exchange can perform internal analyses, as needed. Additionally, they can collectively agree to supply elements in order to answer questions in the aggregate (for example, data that might pertain to the level of preparedness, i.e., ACT/SAT scores).
- An advantage not to be neglected: MOAs based on this design are signed!

**Cons:**

- Inadequate for transactional use: data are entirely retrospective and no information is available for students not included in original cohort(s).
- Sequential process for exchanging data is time-consuming and costly.
- New research questions may require specifications of new cohort(s), which in turn requires agreement among all entities to pursue those research questions and incur the expenses for reconfiguring the match based on the new specifications.
- This design probably does not lend itself to being an on-demand resource to single states or state agencies; any exchange operations likely must be authorized and conducted collectively among participating states or state agencies. Once built, the infrastructure exists for states to specify their own cohorts to satisfy their own unique analytical interests. But, several barriers stand in the way of this design being employed that way, namely: costs may be prohibitive for a single state to take on; agencies may be legally required to identify the public interest in responding to the exchange’s request for data to answer another state’s particular question, which could be difficult if a narrow, state-specific standard for interpreting the public interest is applied; and states may not be willing to contribute the resources necessary to prepare the data extracts another state needs.
- This design could result in laborious and costly repetition given the wrong set of circumstances. States interested in updating data for the original cohort with additional or updated information at a later time would have to re-do the matchmaking process nearly from the beginning. Additionally, if participating states wanted to retroactively incorporate data from additional states based on the same original cohort, to be complete, all states would face the burden of additional matching and extracting.

---

3 This latter example is unlikely to be feasible at the current time. FERPA restrictions on disclosures being authorized only for evaluation and research on educational programs would be one barrier, but a more pervasive one could be that labor market information agencies cite state law and the Privacy Act as prohibiting them from supplying SSNs to any exchange.
2. Fully developed relational database

Using this design, participating state agencies would simply make their data available to other participating state agencies through a centrally managed, relational database. Each state agency would be required to periodically upload a frozen set of data elements that had previously been agreed to collectively by all participating state agencies. Contents would presumably include all students in postsecondary institutions at the point at which data extracts were frozen, plus some subset of all K-12 students (perhaps the students in grades 9-12). In essence, this database would contain a narrower set of data elements but otherwise would basically mirror in contents what the state agencies themselves maintain. The central database would manage identify resolution and the matching process, and would assign a unique exchange identifier to all records. Once all education records were suitably reconciled in the database, the entity given responsibility for database management would then request labor market data elements from participating workforce agencies for just those individuals contained in the database. To do an analysis, a state agency would submit a request to the database management group for matched data for a set of individuals specified by the state agency.

Pros:

- This design would be able to supply more data for multiple years and students; it would not be restricted to a single cohort of students. As such, it would reflect more fully the wide range of activity generated by state investments in education, the labor market and potentially economic development.
- Many of the other advantages applicable to the cohort-based approach apply here as well, including the ability of states to specify which data elements they would be interested in exchanging, states can define cohorts themselves, and research questions that can be addressed by such a database are only bounded by which data elements states agree to exchange within it.
- The design would lend itself easily to producing deidentified data files for approved projects proposed by external researchers.
- This design could potentially be an on-demand resource for participating state agencies. States could query the database for the records of individuals meeting different cohort selection criteria specifically defined to address their research projects. The resulting output for the states might include enhanced personally identifiable data for individuals a state agency could legitimately claim as its own. Or, in the same way the database could provide deidentified files for researchers, it might consist of a deidentified dataset containing information on a broader group of individuals. Such a resource would be useful for answering research questions as they arise, without necessarily requiring the approval of all participating states.
Cons:
- This design may simply be too big and unwieldy to be feasible. Costs could be prohibitively high. And the burden on state agencies to provide data and perform data quality assurance checks would be substantial.
- State agencies may be reluctant to agree to operate such a central database, in part because governance processes and procedures to arrive at such an agreement would be complex.
- Management of the agreement and the database could be complicated and costly.

3. “Plug-in” to National Student Clearinghouse (NSC) core collection

The National Student Clearinghouse (NSC) already collects and maintains a database encompassing many years of student-unit-level data for postsecondary students; that database provides nearly universal coverage of postsecondary enrollments and has expanded to include basic enrollment information on secondary students, provides transcript information, and is considering collecting more course credit information. The third architecture design envisions an exchange that capitalizes on that resource more directly. NSC would be contracted to populate cohorts based on a set of instructions provided by participating states. NSC could also extract corresponding enrollment and awards data. Linking with workforce data would still require a Social Security Number for each student, which would be tricky. For NSC to supply student SSNs they do have, there are two potential requirements: the adoption of new regulations or new guidance on FERPA that would permit disclosure of SSNs without individual consent for the limited purpose of matching to workforce records, and NSC’s willingness to supply SSNs to an exchange without consent of those institutions providing original data. A second possible solution might be for NSC to provide sufficient identifying information to allow appropriate state agencies determine student identity and then supply the SSNs to NSC from institutional records.

Pros:
- This design attempts to take full advantage of existing resources, potentially saving money and time.
- It would reduce the amount of time a state needs to spend extracting and uploading data for the exchange, data that, by and large, NSC already has from institutions.
- This design is potentially more flexible, as states can redefine cohorts to answer additional research questions as needed, without the need for the time-consuming sequential process outlined in the first design described above.
- Highly scalable; adding additional states simply requires NSC supplementing extracts with their data as well.

Cons:
- At least at the present time, the likelihood of finding a legal avenue to allow essential workforce data to be incorporated into this design is remote. Loosening restrictions on the use of SSNs is not probable. Furthermore, NSC has been reluctant to provide data that has not originated within participating state agencies. NSC supplies data only for students a state (or institution) has already identified and has a legitimate interest in, not the reverse
case where NSC essentially declares which students its records show a state or institution should be interested in. The primary reason is that NSC cannot supply these data and remain FERPA compliant without making the implausible assumption that NSC’s roster and a state’s roster of publicly educated students are identical. Finally, redisclosure of personally identifiable information files is impossible as well.

- Cohort definition and creation is out of participating state’s control.
- This database design makes limited data elements available. NSC captures most of the most commonly needed data elements in its main collection, but it does not have important ones of interest to the data exchange, like credit hours attempted and credit hours earned (by term or cumulative).
- Relative to the postsecondary enrollment data, NSC’s coverage is much less extensive at the present time for postsecondary awards (especially certificates) and secondary enrollment and awards.

4. WRIS doppelganger

This design is an attempt to adapt the successful Wage Record Interchange System (WRIS) solution as the architectural model for the data exchange. Under a data-sharing agreement reached with all 50 states, WRIS provides individual state workforce agencies with access to wage records contained in all other states by providing what is essentially a crosswalk of SSNs by states. The appropriate agency in a state can query this table for out-of-state employment records of individuals for whom the querying state has a valid SSN in its own records. Upon requesting information for a set of SSNs, the WRIS clearinghouse queries the crosswalk database for the state(s) where each individual’s records exist, requests those records from each state, and compiles all records into one dataset and returns it to the querying state. The WRIS agreement contains restrictions for how its data can be used and who can access it. But states can submit queries to WRIS at any time for performance-based reporting. The agreement does require each state whose data are used to give express, voluntary approval to the use of those data for research use. WRIS contains data for eight quarters.

The original WRIS only included standard data elements that are captured in the UI wage record files. But some state labor departments have recognized the utility of reporting on how participants in education and training programs, including those offered by colleges and universities, perform in the labor market no matter in which state they are employed. These agencies have developed a WRIS2, which allows the flow of aggregated UI wage record data to be shared with state educational institutions and other education and training providers. Both the WRIS and the WRIS2 agreements are clear in that only a single agency in each state – the one designated by the governor as the Performance Accountability and Customer Information Agency (PACIA), often the state’s workforce information agency – may have individual-level data from the databases in any form. Thus, as promising as WRIS may be for obtaining labor data from all 50 states to incorporate into the data exchange project, currently the agreement expressly prohibits such use, except in the aggregate.
Yet the WRIS experience does offer a potentially flexible architecture to the problem of how to create a linked longitudinal data exchange spanning K-12, postsecondary, and workforce. The core of such a design would be a central repository with elements corresponding to the identifying field used by each participating state or state agency for the same individuals. Each state education agency would upload data corresponding to students’ identifying characteristics to the central crosswalk table. These elements could include first name, last name, middle name (or initial), date of birth, social security number, and whatever other elements the participating groups could agree on for the use of matching individuals. State agencies would also include any agency identification numbers in use for that individual. The operator of the data exchange would use this information to match individuals across states (and possibly state agencies within states) and produce a crosswalk table of identifiers, as illustrated below. The central core of such a crosswalk table would be four columns containing available statewide identifiers in use for each individual, but for illustration, some fictional names (and identifiers) are also provided. If a state had more than one identifier (i.e., only agency-specific identifiers existed), then the table would include more columns of data.

<table>
<thead>
<tr>
<th>Student Name</th>
<th>Hawaii</th>
<th>Idaho</th>
<th>Oregon</th>
<th>Washington</th>
<th>SSN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joe Smith</td>
<td>0473Q</td>
<td>JS9927</td>
<td></td>
<td></td>
<td>335-81-2004</td>
</tr>
<tr>
<td>Jane Doe</td>
<td></td>
<td></td>
<td>00019823</td>
<td></td>
<td>611-29-4334</td>
</tr>
<tr>
<td>Brian Prescott</td>
<td></td>
<td></td>
<td>25009317</td>
<td>B-8440311</td>
<td>298-04-0389</td>
</tr>
<tr>
<td>Peter Ewell</td>
<td></td>
<td>PE7156</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Karen Paulson</td>
<td>0999G</td>
<td>KP4415</td>
<td>41008097</td>
<td>R-0406380</td>
<td>555-33-1111</td>
</tr>
<tr>
<td>Hans L’Orange</td>
<td></td>
<td></td>
<td></td>
<td>L-8656557</td>
<td>402-67-8818</td>
</tr>
</tbody>
</table>

Thus, Joe Smith was at one point a student in Hawaii and Idaho, Jane Doe only attended schools in Oregon, Brian Prescott attended in Oregon and Washington, and so forth. Once a querying state knew where to locate a student’s records, they would request pertinent information from the other state(s), either through an automated process like WRIS, or through a manual process in which the other states would extract and provide the requested data. Let’s say Hawaii wants to know what happened to Karen Paulson after she seemingly dropped out of high school. Hawaii queries the crosswalk table and discovers that Karen has taken a meandering educational path through all four states. Armed with her identifiers from those states, Hawaii is able to retrieve records from the other three states. Using the SSN, Hawaii may be able to request wage records for Karen as well.

Pros:
- This design may represent the ultimate in flexibility and adaptability. If the data-sharing agreement were written broadly enough, states could potentially query the crosswalk table and obtain data as needed.
- Disaggregating the data can potentially be done on the fly, with more frequent updates than would be possible through other means. The crosswalk table potentially incorporates all the contents of a state data system, frozen periodically, so no students would be lost due to an *a priori* cohort specification. The design is not dependent on a research design. New research
questions can be proposed and addressed, although the WRIS model does require research to be approved by participating states.

- Storing and making available the crosswalk data would be relatively straightforward technically. A more limited set of data is required to be held centrally, but there may be many more records sent through any identity resolution process.

- States may not be required to adjust any of the contents of their data systems. In other words, the WRIS doppelganger would return data files with the elements as stored in each state from which data would be drawn, without the necessity of a process to collectively define the data elements to be exchanged. This could simultaneously allow for more elements to be requested by the querying state, at the cost of the data elements received not being in agreement across states. For example, our current MOA does not include numerous data elements that may be of interest. But if a state was particularly interested in assessing students’ academic preparation, they might seek out individuals’ placement test scores obtained in other states, which if stored in state data systems may not agree as to the assessment itself, relevant cut scores, value labels, etc. The WRIS doppelganger design might provide a way for states to exchange such data, but the requesting states would need to take extra care to understand and appropriately interpret and use the data elements it would receive from other states through its query.

- States would also have the flexibility to determine which data could be shared through this process with its peers.

**Cons:**

- This design could incur high development costs, or additional ongoing costs depending on how often the system is accessed. Creating the automated capability that WRIS has for gaining access to other states’ data is a significant development project, one with uncertain prospects concerning FERPA compliance. Without an automated system, states could potentially be asked repeatedly to extract data for another state’s query of the system.

- Without a formal process designed to commonly define data elements and derivations as a necessary part of the data exchange’s operation, data element definitions may not agree, or some states may restrict access to data elements on the basis of inconsistent interpretations of what may be necessary for research and evaluation. (Adherence to CEDS would help alleviate some of this as a potential pitfall.)

- States would need to devote time and energy to determining how to provide aggregate reporting for a region. This architecture design does not eliminate the need for a central entity to collect, clean, and make available a de-identified data set for a multi-state region, and it is unclear whether or not states would be willing to pay the added costs for such a feature.

- It is unclear whether a crosswalk table such as this could provide any avenue into the workforce information. The most straightforward way might be to include SSNs in the crosswalk table operating under broad agreement including workforce agencies. Doing so might allow the participating states to access UI wage record files themselves. This may be stretching the bounds of FERPA and possibly the governing statutes and regulations on the workforce agencies’ data as well. So it seems probable that the process would still require
the intervention of state educational entities in requesting wage records from multiple states for a given set of SSNs. The current WRIS agreements do not appear to provide space for using them to obtain individual-level data, even from participating states and despite the fact that this design adopts a similar architecture. To gain access to workforce data for all individuals in a multistate region, there probably still needs to be an entity compiling all related records in order to submit them to state labor agencies to gather UI wage records.

- The WRIS model only maintains data for a short two-year timeframe. Adapting it for this use would require more years of data to allow for meaningful analysis.

Having presented these four broad design concepts and some associated tradeoffs, the question is which of these offers the best chance at a sustainable resource for tracking human capital beyond the conclusion of the grant. There may be other possibilities, or mixtures of elements within these four, as well. But specifically, how do we think about the relative merits and demerits of these approaches for long-term planning?