**Recommendation Paper**

*Early Childhood Longitudinal Data System*

*Cross-Departmental Project Team:*

*WI Department of Children and Families*

*WI Department of Health Services*

*WI Department of Public Instruction*

*WI Department of Workforce Development*

**Entity Resolution Tool,**

**Potential System Architecture,**

**Stakeholder Involvement**

**and Sustainability.**

*12/17/2012*

**2012/2013 Proposed Considerations/Recommendations for discussion by EC-LDS Project Team; submitted by Carol Noddings Eichinger and June Fox**

**In reviewing the Project Charter and the current recommendation papers submitted for the Unique Identifier Committee and the Identifying Capacity Committee, it appeared useful to compile some history and thoughts about several other key deliverables identified in the Charter for the Feasibility Study; Unique Identifier Question (Entity Resolution Tool), System Architecture, Stakeholder Involvement and Sustainability.**

* **ENTITY RESOLUTION TOOL**
	+ The State Support Team provided guidance during discussions for the federal LDS site visit and during conference calls with Jeff Sellers through Missy Cochenour.
	+ These discussions led to consideration of the open source tool Oyster/Arkansas . Some members of the Project Team had also viewed this tool as demonstrated at the SLDS Best Practices Conference in November of 2011.
	+ In spring 2012, several Webinars provided by Arkansas demonstrating OYSTER included IT personnel from DHS, DPI and DCF.
	+ The Unique Identifier workgroup studied the entity resolution solution: OYSTER. The **O**pen-S**yst**em **E**ntity **R**esolution (or OYSTER) is an effective proven open-source (free) product that matches to determine entity resolution. The more recent version of OYSTER makes use of Knowledgebase Identity Management (KIM)and separates out information, as shown in the diagram below.
		- The importance of privacy and security is well illustrated below. KIM contains the personally identifiable information (PII) which is held separately from the de-identified data (TrustED). TIM contains and manages the crosswalk information. This separation is well-vetted as a reliable approach to manage a high-confidence rate of accurate identification of individuals, as well as providing tight security and privacy of data and identifiers.
			* **Definitions for terms below:**
			* PII – Personally Identifiable Information
			* KIM – Knowledgebase Identity Manager
			* TIM – TrustEd Identifier Manager



* + The Division of Libraries and Technologies (DLT) at DPI is investigating an open source entity resolution tool from Idaho called Education Unique Identifier (EDUID ), which was introduced to DLT at the MIS conference in San Diego in February, 2012. This tool also allows assignment of a unique ID that travels with the child from school entry through the workforce, later to potentially become an educator ID for teachers. A high level Webinar of this tool was presented by Idaho to IT and other interested personnel from DPI, DCF and DHS. The EC LDS project will benefit from the information obtained by the current analysis being performed by DLT.
	+ DPI has in place a Webservice used for sharing of post-secondary student information from K-12 LDS across the University of Wisconsin System, the Wisconsin Association of Independent Colleges and Universities, and the Wisconsin Technical College System .

**Recommendation: Year one of the Race To The Top Grant will involve data matching experts in selection and implementation of an open source entity resolution product.**

* **POTENTIAL SYSTEM ARCHITECTURE: Federated vs Centralized**
	+ Please refer to the diagram and chart below for a presentation of differences between federated and centralized SLDS, provided by the Federal State Support Team (also see enlarged diagram at end of document if you want a larger version).



Federated vs. Centralized Data Warehousing

|  |  |  |
| --- | --- | --- |
|  | Federated | Centralized |
| Data Ownership | Data ownership is with the source agency with no need for shared data stewardship. | Data ownership is with the source agency with shared data stewardship with the centralized data warehouse agency/entity. Responsibility for this data stewardship should be spelled out in memoranda of understanding (MOU). |
| Staff resources | Staff resources are required of each source system to oversee and maintain required data access. In addition, support will need to be given to the extract, transform and load (ETL) processes to reflect changes in source data systems and data element modifications. | Staff resources are required of each source system to oversee and maintain required data access. In addition, support will need to be given to the extract, transform and load (ETL) processes to reflect changes in source data systems and data element modifications. Staff will also be needed to support the centralized data base system. |
| Technical Requirements | Each source system will need the required hardware and network bandwidth to facilitate and process external queries (ETL tools), conduct matching processes and returning the resulting dataset. There will also be a need to deliver the matched resulting dataset, i.e. portal or business intelligence (BI) solution. | Each source system will need to be willing to allow access or provide the data to be included in the centralized data system. An infrastructure to support the centralized system along with ETL tools, conduct matching processes and storing the results. There will also be a need to deliver the matched resulting dataset, i.e. portal or business intelligence (BI) solution. |
| System Performance | Subject to longer delays in data delivery due to load on source systems, etc. Agency specific performance issues can affect the performance of the entire system. Also the possibility of limited or narrow windows of processing time due to other/competing priorities.Relatively new technology; accounts for less than 10 percent of all data warehouse projects; not a proven technology. | Data extraction is generally fast since all data matches have occurred in the transformation and load steps. Match once, use many times. Scheduled extracts can occur on source systems during off-peak hours to minimize impact on sources. Centralized data system architecture can be designed specifically for this purpose, thus increasing response times.Established technology and procedures; proven technology. |
| Privacy/Security | Primary responsibility is with the source system agencies. Secure process needed for handling of data queries | Primary responsibility is with the centralized data system agency/entity as the data steward, but is dictated by source system agencies via memoranda of understanding. Security is handled through access rules for users. |
| Data updates/corrections | Data resides within each agency. Each agency is responsible for communicating and possibly updating the data extract processes to reflect changes, corrections or updates.  | Establish process for ETL either when data are changed (if required to have near real-time data in centralized data system) or at a specific periodicity to capture changes, corrections or updates.  |
| Data Availability | Based on when data are available in the source and made available for extract. Access to data is determined by source agency. | Based on when data are available in the source and made available for extract. Access to data is determined by source agency via MOU. |
| Data Quality | Dependent on processes implemented at each agency | Process for data cleansing apply to all data as agreed upon by the source system agencies; consistency of data cleansing processes and data quality checks. |
| Implementation | Generally requires less time; although equal time is needed to determine requirements and processes for ETL and data provision. | Longer implementation period due to the need to build the centralized data system database/warehouse. But equal time is also needed to determine requirements and processes for ETL and data provision. |
| Expanding to include other agencies | The addition of any required hardware and other resources (as mentioned above) required for data queries/matches across the system. Writing ETL processes and matching/integration rules. | Potentially supplementing or expanding centralized data system architecture to accommodate additional agency source system data. Writing ETL processes and matching/integration rules.  |
| Production of standard reports | Dependent on an agency accepting this as a responsibility. | Can be an automated process; less expensive and timelier to accomplish. |
| Sustainability | Possible approaches are for each participating agency make their contribution for the corporate support of the processes needed for the federated system. This may be a deterrent for agencies to participate. Another approach would be specific appropriation that is allocated to each participating agency, based on a funding formula. | Possible approaches include a state appropriation to the centralized data system agency/entity for the development and ongoing support and maintenance of the centralized system. This would have no fiscal impact on the participating agencies. Another approach would be for each participating agency to pay for a proportional part of the needed funds for the support of the centralized system, in a cost recovery model. This could be a deterrent for agencies to participate. |

* + Throughout our project workgroup discussions during this planning year on various topics (Unique Identifiers, Data Governance, etc.) workgroup members have touched on the topic of the Federated vs Centralized model for our basic System Architecture.
	+ In analyzing the information provided to us above from the State Support Team, combined with information presented at the Washington DC SLDS Best Practices Conference (October, 2012), information provided by the Washington DC Intelligence For Social Policy Fall Conference (November, 2011) and statements of strong preference expressed by the cross-departmental team members, current thinking moves us towards a federated or hybrid federated model rather than a centralized model.

**Recommendation: The proposed System Architecture for the EC LDS is a Federated Hybrid Model. We currently envision a standard Federated model for entity resolution and the linking of data across departments. The data will not be stored centrally, but cached and returned to the approved requester for their use/analysis/reporting purposes. The use/analysis/reporting can take place in various analysis/presentation layer tools, as appropriate to the department making the request for data. The nuance which makes our current recommendation a Hybrid version concerns the choice (still to be made) of a common presentation layer for standard shared reporting/shared dashboards. MOUs, security, privacy, etc. around the various presentation/analysis tools/data/user access will still need to be decided and agreed upon during the Build/Test/Implement phases of the upcoming EC LDS project.**

* **PRESENTATION LAYER**
	+ DPI recently launched the WISEdash presentation layer for the current LDS. DHS and DCF currently use Business Objects as a presentation/analysis tool in-house for department specific analysis of data.

**Recommendation: The Presentation Layer** **will be a later intersecting focus during the Design/Build/Implement phases of the upcoming EC LDS project.**

* **STAKEHOLDER INVOLVEMENT**
	+ Communication Plan developed early in project itemizes the important processes for keeping potential stakeholders involved, including the following activities:
		- Department Briefings held in Fall 2011 and Winter 2012
		- Project Team meetings monthly
		- ECAC meetings quarterly
		- Data Roundtable “Business Requirements” February 2012 session with 72 participants
		- IT discussions in each department ongoing
		- Monthly participation in LDS Steering Committee (DPI)
		- Monthly status reports
		- Monthly website updates
		- Ongoing contact with local initiatives
			* Brown County LDS Initiative
			* Racine County LDS Initiative
			* Milwaukee Public Schools LDS Initiative
			* Chippewa Falls LDS Initiative
	+ Participation in the PTAC workshop in Utah in August allowed a team of Project Team members from DHS, DPI and DCF to explore national guidance on creating MOU and data governance structures for sharing early childhood data across systems. An initial MOU is being crafted for the proof-of-concept data sharing anticipated in early 2013, which will lead to a more robust ongoing MOU process that will facilitate the ongoing sharing of data across the three departments. Appropriate legal counsel will be sought to ensure all privacy and security regulations are carefully addressed within the MOU.
	+ In addition to the identified activities above, the MOU process between the three departments will further engage management personnel from program areas and at the Secretary/Superintendent level in agreeing to the EC-LDS intentions.
	+ The update of website materials and continued correspondence with expanding pool of stakeholders, including the initial Data Roundtable attendees will continue.
	+ During the formation of Data Governance structures, members will be identified and become leading champions of the intentions of data sharing across the three systems.
	+ As the EC-LDS formal data sharing architecture is operationalized, a core group of data experts from around the state will convene for in-person planning & development meetings for training modules on using data to inform decisions, as well as dashboard & guided analysis workbook planning and design sessions .
	+ From this development of training materials, the training provided at the department level in both accessing, interpreting and using the data will emphasize the “value-added” in sharing data and receiving data from other program areas, further encouraging stakeholder participation and appreciation.
	+ As training rolls out to the local agencies and districts, stakeholder participation and appreciation at those levels will also solidify.
	+ It is anticipated that the core set of questions the data may answer will evolve and grow as stakeholders explore the further potential uses of the available data.
	+ The stakeholder pool with continue to evolve as the EC LDS matures. For example, future plans include linkages with workforce development data and stakeholders. This will complete the P20W picture.
	+ It would be important to convene a key group of EC Researchers to consider initial direction for the department specific research analysts (as defined in our Race To The Top Grant Application) and to intersect with DPI’s Research Steering Committee and any similar research committees located at DHS and DCF. Include VARC and IRP. Begin formulation of research agenda, with the emphasis being on cross- departmental data sharing.

**Recommendation: The above bullet points illustrate our current plan for Stakeholder Involvement. This plan is being executed within the current Feasibility Study (planning phase) project and will continue during the Design/Build//Implement Phases of the upcoming EC LDS project.**

* **EC LDS SUSTAINABILITY**
	+ With the receipt of the Race To The Top – Early Learning Challenge Grant one project, within the portfolio of 9 projects, is the Selection and Implement of an Entity Resolution Tool (Matching Tool) to occur during 2013. The testing of this selected tool will jump start the topic through establishing an MOU and “trying out” the ability to match and identify/locate children (unduplicated counts) across the three agencies.
	+ The “value added” by the testing of the Matching Tool above with initial access to data will also increase demand for the additional data answered by the additional data sets. It will be important to articulate the key questions, program areas and data sets, and data elements to establish the initial data components and decision-making potential at the department level first via the Sustainable Data Governance Committees.
	+ Engaging local agency and district personnel early to influence the roll out of access to the local data will again increase demand and emphasize “value added”.
	+ In the future, the three departments could consider exploring the options of potentially braiding funding to continue to support the data sharing in an ongoing fashion, as well as the possibility of legislation to further define the agreement. Continued research on legislation passed with parallel projects in other states could continue with guidance from the Federal State Support Team (SST).
* **Recommendation: The above bullet points illustrate our current plan for EC LDS Sustainability. This plan is being executed within the current Feasibility Study (planning phase) project and will continue during the Design/Build/Implement Phases of the upcoming EC LDS project.**

