

A Comprehensive pK-12 Data Model

Project Executive Summary

Phase One:
A Guiding and Stakeholder Convening Draft

October 2005

A Comprehensive pK-12 Education Data Model

I. Project Rationale

The increase in the use of accountability models and diagnostic analyses to improve student achievement to fulfill extensive reporting requirements is leading districts and states to design, develop, implement, and apply sophisticated data systems. The architecture and content of these data systems define the breadth and depth of research and diagnostic analyses each system is able to support. The content of state data systems is dependent on the data collected by district data systems, which, thereby affects the state's ability to aggregate and use data for analysis and planning. The content and architecture of state data systems, in turn, limit the systems' analytical capabilities and, therefore, their ability to support instructional improvement and help close student achievement gaps. The content and structure of data systems also affect the states' ability to share data on students who move across state boundaries.

The only way to meet today's accountability demands is through the development of a comprehensive and dynamic common pK-12 data model. To date, the closest access to data models that schools and states have is in proprietary models developed by vendors and implemented in their software applications. With the majority of districts (more than 10,000) in the US having fewer than 2,500 students, many districts cannot afford these data solutions or they must enter into consortium purchases which do not allow them to tailor their data utilization to the needs of their educational stakeholders. Most states and school districts cannot make the necessary financial investment or do not have the technical expertise to develop such comprehensive data models.

This project was conducted to determine how best to provide guidance and best practices for states and districts putting together their enterprise data structures. The National Center for Educational Statistics (NCES) asked to identify the need and scope of work for a possible follow-on project to generate a comprehensive pk-12 education data model. The resulting report discusses the need for a comprehensive data model and provides a first attempt at categorizing how pK-12 data will be sorted and represented in the NCES state and district data models - or a high level conceptualization of the data model. To accomplish this, a series of activities were conducted:

- An advisory committee made up of state and local data systems administrators, SIFA data experts, and others knowledgeable about state and local education data met on two occasions. They worked virtually to discuss topics such as target audiences both for this document and for a comprehensive education data model, to provide guidance on model reviews, and to provide input into the recommendations contained in this document.
- Sample data models were reviewed to identify essential components and initial indications of data relationships. This "gap analysis" provided input into what components should be called, helped to determine what the ideal product should look like, and justified the need for a comprehensive education data model. "Data models" include in this analysis included national (SIF, PBDMI/EDED, NCES Handbook, NCES Common Core Data), state (California, Virginia, Maine) and local (Washington DC and Richmond County Schools)

A Comprehensive pK-12 Education Data Model

II. pK-12 Data Models

What is a Data Model?

A data model is a conceptual representation of the data structures (tables) that are used in expressing and communicating business requirements. Done correctly, models visually represent the nature of data, the business rules governing the data, and how the data will be organized in a database. The goal of a data model is to make sure that all provided data objects are completely and accurately represented. A data model can also be used to communicate within the business and across business domains - as seen in education.

A data model also may provide a method of visualizing the informational needs of a system. It typically takes the form of an ERD (Entity Relationship Diagram) and almost always must include a data dictionary of some kind to be complete. Three basic types of data models are used:

Conceptual Data Model (CDM) - A CDM represents the overall logical structure of the sets of data, independent of any software or data storage structure. A conceptual model usually contains data objects not yet implemented in the physical databases. It gives a formal representation and factoring of the data sets needed to run an enterprise or a business activity. Conceptual models tend to lend themselves to a higher level stakeholder audience such as administrators. This type of model provides a “high level” overview of the flow of data without getting into the granular details. This model also paints the “big picture” of the data needs, in this case across the pK-12 realm.

Logical Data Model (LDM) - A LDM fills in the gap between a conceptual and physical data model (PDM). CDMs are completely devoid of database-level information while PDMs are specific to a certain database management system (DBMS). A LDM stores those characteristics without adding anything specific to a single DBMS. Logical data models are more granular in detail and are useful to audiences who want to understand the relationships between the data elements and the desired flow of data. While this type of model still captures the conceptual outline of the data, the logical or expected flow of data is exposed in greater details. This type of model is typically used by data administrators and managers.

Physical Data Model (PDM) - The PDM specifies the physical implementation of the database. With the PDM, one considers the details of actual physical implementation and the specific data elements correlated to fields, dimensions or facts in the actual data base. It takes into account both software or data storage structures. People often modify the PDM to suit their performance or physical constraints. The physical data model is one that is the hardest to display and maintain. This type of model is often used by the software developer and technology audience. It is ever-changing and evolving as systems are added and removed from the model. This type of model is said to “never be finished.”

Past Attempts at Data Modeling

While there have been efforts to describe the various educational data relationships, these models tended to be narrowly focused at the local (report generation), state (reporting for accountability/funding) and national (PBDMI/EDEN) levels. Each effort has focused on a particular set of data identified to fulfill some mandated/operational task. The majority of work done around the development of comprehensive pK-12 data models can be seen within

A Comprehensive pK-12 Education Data Model

the vendor community in the development of data warehouses and/or student information systems to be sold at the local or state level. These product lines have been comprehensively developed, although most do not have a clear understanding of educational business cases and relationships. They also receive ongoing maintenance to promote company proprietary business advantages.

This project was designed to gather examples of past efforts, develop a high level representation of the extent of work, and provide a “gaps and overlap” analysis of data currently utilized in the pK-12 space. This representation will include various national and state “data models” which takes into account representations of the data needed for education planning, operations, and reporting.

There is currently no comprehensive data model at the national level. There are numerous data dictionaries for various audiences: the National Center for Education Statistics (NCES) data handbooks, the USED PBDMI/EDEN data structures, and the Schools Interoperability Framework Association (SIFA) data model. They, however, provide little guidance on the relationships and interconnectedness between those structures and data sets. At the state and district level almost all models are physical models and representations of their database structure.

III. Project Findings

Summary of Data Model Comparisons

There are several driving conditions influencing the comparisons:

1. Most of the models were, or are close to being, physical data models. While this is appropriate for seeing what is actually being implemented at the state, district and local level, the particular database management systems and organization-specific workflow designs make it difficult to abstract the conceptual principals that gave rise to the model.
2. It was discovered that many of the organizations never built a conceptual data model at all. They had their technologists build a logical data model with the physical model in mind.
3. Some of the “data models” are not relational models at all. They do not address the relationships between the elements; rather they are lists that include the narrative definition and the code values and possibly the physical field description of the elements. Essentially they are data dictionaries.
4. Due to the wide variety in the models and differences in scale, we compared the models at the level of section and category so as to have an “apples to apples” comparison.

Upon close comparison, the existing models varied greatly in design and function. While many were physical in nature, it was possible to draw from them the essence of data they were attempting to represent. While the NCES Handbooks represented the most descriptive data, it is not in the true form of a data model. Although it represents the most comprehensive elements for education currently available, the PBDMI / EDEN information is very comprehensive at the federal level but data is not found to be included in the SEA or LEA models. This suggests that the data elements at these levels would not be included if a comprehensive data model was attempted by an SEA or LEA. If they were included, this would streamline reporting and reduce the burden on schools and states.

A Comprehensive pK-12 Education Data Model

While overlaps have been exposed, it is obvious the data represented by all the models varied significantly at all levels. The greatest amount of overlap was in the enrollment, program and assessment sections. One would think that the student and staff domains would be well represented in all of the models but, comparison of the overlaps revealed that this was not the case.

The largest gap that the expert panel encountered was the lack of details at the class level. They felt this gap has the biggest impact on education and learning. It exposed the data needs that are inherent if LEA and SEAs are to do precise and comprehensive longitudinal data studies. However, this level of detail would be a focus area that should be addressed in a follow-on activity. The panel recognized another gap as the need for relationships between the data and domains to be in a conceptual and logical manner. Without this detail in a comprehensive data model LEAs and SEAs may struggle to understand the data at its lowest level where it impacts students and staff.

One of the conclusions of Phase I of this project is also one of its assumed starting points: most districts do not have the money, technical resources, or time to build a fully-developed data model. Instead they build their data structures piecemeal out of what they get from each vendor or as each project is completed. Thus there is little or no uniformity or standardization, upgrades become difficult to complete and solutions must include multiple vendors with different requirements. While that situation may not change, the presence of a definitive comprehensive example of a data model, can serve as a starting point to significantly improve the design, implementation, and utility of education data systems.

IV. Conclusion

It is clear from the work done by SIFA and the expert panel that there is a real need for a comprehensive pK-12 education data model. It would also be critical that this model is kept up to date through well-defined periodic review and stakeholder input mechanisms. Such a model will help to ensure that data systems can meet the needs of educators, not the requirements of vendors. Educators are the essential reviewers and users of a data model. If the questions that need to be answered cannot be obtained by the relationships of the data, then the model or system is flawed. Vendors also benefit from the use of such a model because it will indicate important components for educators.

If a comprehensive data model is developed with the information gleaned from this report, another key component to its usage and implementation will be the documentation and dissemination of best practices usage and possible tools for model utilization. A true “best practice” data model should show how all essential questions can be answered through the use of data within the education system. This could be considered Phase Three in the successful development of this comprehensive pK-12 data model.

“In the end, the real line of accountability will be drawn when a district goes to use or implement the comprehensive data model” – a sentiment delivered by a member of the expert advisory panel.