Texas
Student
Data
System

TEXAS STUDENT DATA SYSTEM

Lessons Learned: Prototype Release 1 Development & Technical Notes

Zeynep Young, Double Line Partners
Sharon Reddehase, Double Line Partners
Ed Comer, Cooper Consulting Co.

This document and the project described herein was led by the Texas Education Agency and supported by the Michael & Susan Dell Foundation. The project was conducted by Double Line Partners, LLC.

The experiences and conclusions related in this document reflect those of the authors and do not necessarily reflect those of the Texas Education Agency or the Michael & Susan Dell Foundation.

Table of Contents

Document Purpose

Performance management and the effective use of data to inform instruction are at the forefront of education reform. Many education organizations have developed successful systems that collect and report data from disparate source systems through a set of user-friendly dashboard tools.

Leveraging the knowledge of others and building a set of tools that incorporate best practices allows local education agencies (LEAs), state education agencies (SEAs) and related education organizations to replicate this work more cost effectively and efficiently. This, in turn, creates an environment of continuous improvement that ultimately has the potential for improving student outcomes on a wider scale than what was possible in the past.

Careful study of the lessons published from others' projects helped this team avoid common pitfalls. In that spirit, this document outlines the experience and lessons learned as Double Line Partners and Cooper Consulting, in collaboration with Lubbock Independent School District (ISD), implemented the initial prototype of the District Connections Database project.

Below is a discussion of the development approaches used in the initial prototype, with highlights of the technical team's encountered challenges. Although this document will be of interest primarily to a technical audience, business users also can benefit by understanding some of the scope, timing and resource issues discussed throughout.

There will be two additional releases of the prototype through June of 2011. Similar documents will be developed to reflect experiences and observations from those prototypes.

Background

Overview of Texas Student Data System

In 2009, the Texas Education Agency (TEA) launched a five-year effort to redesign and enhance its statewide data system with three primary goals:

1. Build a platform to deliver relevant and actionable data back to educators to continually improve performance(e.g., an early warning system);
2. Alleviate the data collection and submission burden on school districts and improve data quality; and
3. Integrate key data into TEA's P–20 data warehouse to better understand students' preparedness to contribute to the 21st century workforce.

The goals for the new system called the Texas Student Data System (TSDS) were developed through detailed background research and extensive consultation with a wide array of education stakeholders.

TSDS has four key components:

1. **District Connections Database (DCD)** - A new operational data store and data warehouse to ease the burden of compliance data collection and reporting and to populate user-friendly dashboards containing timely, actionable student data
2. **Voluntary State-sponsored Student Information System (SSIS)** - An opt-in, voluntary SIS offering for districts, hosted by the state
3. **Public Education Information Management System (PEIMS)** - A repository for certified data for state and federal compliance and accountability reporting
4. **Texas P–16 Public Education Information Resource (TPEIR)** - A statewide longitudinal data warehouse, linking pre-K, college readiness, higher education and workforce data with K–12 data

Information regarding TSDS can be found at: http://www.TexasStudentDataSystem.org.

Background on DCD Project

In 2009, the Michael & Susan Dell Foundation announced a $10 million grant to support the development and the implementation of the DCD. Double Line Partners, LLC were selected as the technical service provider to develop the dashboards, collect stakeholder feedback, and develop a prototype and documentation of the DCD to be implemented at TEA.

The initial work of the project focused on defining a set of dashboards informed by national research and best practices from over 20 urban school districts across the country. The goal was to define a limited set of critical, student-level metrics that would be predictive of student performance and actionable by educators. Once the dashboards and associated metrics were initially defined, the team moved into the development phase, focusing on the following:

- Developing a data exchange standard for moving data from the various district systems into the DCD
- Creating a mapping and strategy to meet state accountability reporting from the DCD
- Acquire a representative set of source data, map and transform it into the data exchange standard for input into the DCD
- Develop a prototype that moves district data into the DCD and drives a set of student and campus dashboards
- Refine the metric definitions based upon a variety of feedback and lessons learned
- Create a set of documentation for others to benefit from the work

What the first prototype looked to prove

The first release of the DCD prototype aimed to test a number of hypotheses, as follows:

1. The source data exists at the LEA level to drive the dashboard metrics.

2. The accountability data currently supplied by the LEA to the state can drive a useful subset of dashboard metrics.

3. The source data is transportable from district sources to the DCD via an XML-based data standard.

4. The overall DCD system design and data standard “works,” meaning that data can flow from the LEA source system via an XML-based standard to the DCD. Once in the DCD, the same data can generate dashboard metrics and can also generate state accountability data submissions.

What the first prototype did prove

1. There is plenty of data at the LEA level to drive useful metrics. The challenge is to identify, acquire, transform, and unify the data.

2. The state’s accountability data is, in most cases, different from that required to drive the dashboards, though they share a common set of source data. Using the current standards for delivery of accountability data would power only a small subset of metrics. However, using the raw source data proved successful in driving a large set of metrics.

3. The draft CDM XML standard provided highly successful in moving data from the various district sources into the DCD.

4. The prototype proved the validity of the draft CDM XML data standard for flowing data end-to-end through a statewide data system, from raw source through to the dashboards. Once

in the prototype DCD, the system generated both dashboard metrics and accountability data submissions.

Project Team Composition

The project team consisted of a diverse, cross-functional team with staff from the TEA, Lubbock Independent School District, Double Line Partners, Cooper Consulting, and the Michael & Susan Dell Foundation. TEA staff provided oversight, information on the current technical landscape at the agency, and advised on education policy issues around accountability and other data collections. The Lubbock ISD team offered tremendous insight and practical experience with student data and systems. The Lubbock ISD teachers were also instrumental in refining the dashboards to provide maximum value to educators. The Michael & Susan Dell Foundation provided financial support and dedicated foundation staff time for project and technical support. Double Line Partners served as the lead technical service provider and the primary **analysis team**. Cooper Consulting was selected as the **development team**.

The analysis team was responsible for defining and designing the dashboards based on education research findings, best practices from a number of large school districts, and user feedback and focus groups. The analysis team had a core of three people, and expanded to seven people to cover regional stakeholder forums held around the state. The core analysis team members had management consulting experience and had previously been engaged in several district-level performance management projects.

The development team was responsible for developing the software that (a) extracted, transformed and loaded source data into the DCD and (b) structured the data for display in the dashboards and associated drill downs. The development team consisted of a project manager, two business analysts, a database lead, a user interface lead, an infrastructure specialist and a tester. The development team members had over 12 years of experience providing software solutions and consulting to TEA, including being the primary vendors supporting the accountability system since 1999.

Both the Double Line and Cooper teams had significant experience in public education, but from different aspects. The analysis team included experts in education data visualization and decision making. The development team included experts in education data definition, interchange and data warehouses.

Here's what went well

Between the various members of the analysis and development teams, the project benefited from having a robust set of skills to meet the challenges of the project. This was enhanced by the deep level of experience and seniority in the team. Though the members of the team had different skills, every member of the team had significant experience working with education data and visualization. That shared experience, plus already having a shared language to discuss the acronym-heavy details, unified the team. Co-location enabled the team to be highly productive and to cross-fertilize and build upon each other's ideas and experiences.

Here's what went not-so-well

Toward the end of the R1 phase, developing the documentation required to share and validate results stressed the capabilities of the engineering-oriented development team. Adding a technical documentation specialist is planned for R2.

Lessons to carry forward

The project depended heavily on the experience, skills and seniority of the project team. In particular, having a combination of analytical team members who had significant district experience, coupled with technical experts who were familiar with the state's accountability system, proved to be highly valuable.

Drafting the Data Exchange Standards

A major goal – and challenge – of the overall TSDS project, and in particular the DCD effort, is exchanging data between systems. There's no shortage of data standards in the education space: Accountability reports from a district to a state are submitted via a data standard, standardized test scores are sent to districts from vendors in a standard data format, student grades are exchanged between systems at a district using a different format, and so on.

As a starting point in the DCD development process, the team analyzed existing data exchange standards. Although an existing standard was not found that appeared to fulfill all the data requirements, existing data standards and references formed the basis of the eventual data work.

Initially, the team reviewed the National Education Data Model (NEDM).[1] NEDM is a conceptual representation of the PK-12 education data domain. It builds on a number of previous standards efforts, including Education Data Exchange Network (EDEN) and EDFacts record level elements, the National Center for Education Statistics (NCES) Handbooks, among others. However, NEDM is not a standards instrument for data exchange. Rather, as a resource, it provides general data guidelines, depicting the ontology and relationships between large collections of data elements used in PK-12 education.

The Schools Interoperability Framework (SIF) provides both a very detailed education data model and transactional components. The team reviewed the data structures and elements in SIF, and, where feasible, aligned to those data structures. The Post-secondary Electronic Standards Council (PESC) provides several good concrete data exchange schema, and a well-

[1] National Education Data Model, U.S. Department of Education, March 3, 2010, http://nces.sifinfo.org/datamodel/Index.aspx.

thought-out data framework. Though the PESC data structures focus on post-secondary information, the architecture and overall design was a valuable reference.

Beginning with those existing standards as reference points, the team focused on developing an XML-based canonical data model (CDM). Generally speaking, a CDM defines a set of entities and their attributes and relationships relevant to a specific domain. In this case, the CDM domain is defined as P-12 education data complete enough to satisfy accountability reporting and also drive student performance dashboard metrics.

The resulting TSDS CDM is a technical *reference model* for P-12 data that can be used for many data exchange purposes within this context. The reference model was used both to define the TSDS XML standard for data exchange and the logical database model for the DCD – and to keep the data structures in both uses aligned.

As a point of reference, the data defined by the TSDS CDM is very similar to the data in many district-level data warehouse implementations.

Creating the CDM-based XML standard was the most technically contentious aspect of the design activities. Although XML schemas define a well-formed language for data interchange, there are many options for styles or design patterns. After significant technical discussion and analysis, the team selected the following structure for the CDM XML:

- A CDM XML core schema provides a set of domain types easily reused in creating concrete interchange standards.
- A convention is defined to support state-specific or even interchange-specific extensions to the CDM core.
- An interesting mechanism was defined to support interchanges where a lookup must be performed as part of the interchange. For example, when importing test scores, identity information in the interchange must be used to link (in a longitudinal sense) scores to existing student records.

The benefit of the resulting structure is that it provides maximum flexibility in constructing different interchange schemas using reuse and/or extension while staying completely within the semantics of XML.

Here’s what went well

The CDM development benefited from having a specific domain and purpose. The state’s accountability system provided a wide-but-bounded set of data elements, and the student performance metrics provided a lens through which to structure the data elements. The CDM also greatly benefited from being part of a concrete implementation of the DCD and a concrete interchange of LEA data from an ISD. The task of mapping real LEA data refined and validated the draft CDM, particularly with respect to what specific data was important to capture in a source system. The task of creating a DCD database schema based on the CDM served to refine the CDM even further, particularly in the structure of associations.

While pursuing the end objective of an XML-based exchange standard, the team's development of an overarching reference model at first was highly successful. The UML class diagrams served the purpose of understanding and expressing the education data semantics while clearly separating the deliberations over XML structure.

In defining the structure of the CDM XML standard, the team worked hard to avoid a compromise solution that did not fully meet all of the objectives. This persistence led the team through a number of difficult deliberations and to a solution that can serve as a model for other exchange standards in the P-12 education domain.

Here's what went not-so-well

Although excellent XML and data modeling tools exist, the team did not find a tool-set that completely and elegantly fulfilled its needs. In particular:

- The tools were not extensible enough to embed custom metadata that the team identified as being necessary or useful.
- The technical documentation auto-generated by the tools was very good, but not flexible enough to produce the full suite of documentation required for the project.
- The tools were not open enough to link and maintain alignment across the UML and XML and logical SQL models.

Lessons to carry forward

The CDM will go through at least two additional cycles of refinement and validation against a wider set of real LEA data. However, the lessons to date of the CDM for this project have implications for other statewide standards efforts:

- Data exchange standards should be crafted within the context of a specific, well-defined scope. Even though the scope of the TSDS data exchange is wide enough to encompass both a complete set of accountability data and a range of student performance indicators, having clear boundaries made the technical work possible in a relatively short period of time.

- The U.S. Department of Education is coordinating a Common Data Standards Initiative[2] designed to help state and local education organizations define a minimal common set of key data elements with the goal of improving comparability and share-ability across systems. Future versions of the CDM will use that resource as another point of reference. In addition, the elements in the Common Data Standard model are mapped to existing standards like SIF, PESC, NEDM, EDEN, the NCES Data Handbooks, etc. By mapping the TSDS CDM elements to the Common Data Standard elements, the effort can "inherit" the mappings to SIF, PESC, and other standards.

[2] See http://www.CommonDataStandards.org for details.

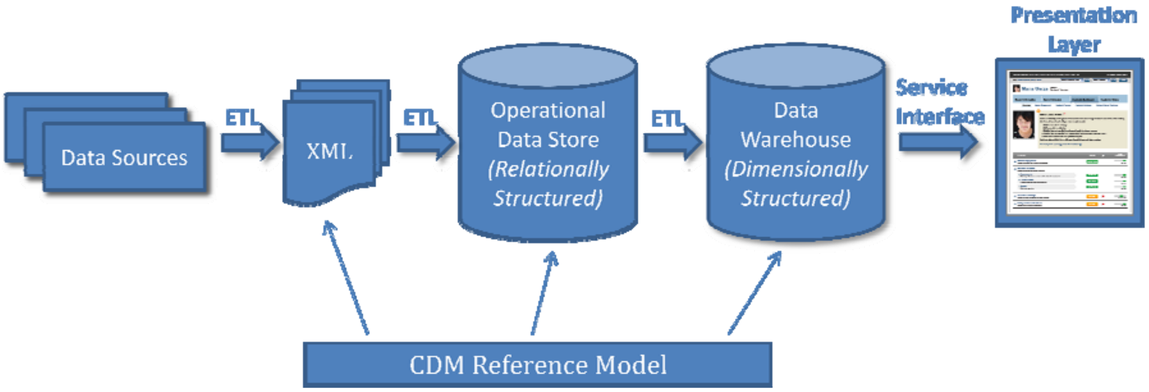
- Although a system can't "implement" an ontological model like NEDM for data exchange, references like NEDM and the new Common Data Standards Initiative have value as a starting point for drafting and validating concrete data exchange standards. NEDM, for example, can serve as a checklist for ensuring completeness of data entities and relationships between entities, and the Common Data Standards Initiative will provide a starting place for the technical structure (e.g., field widths, XML types) of the entities and elements.

- Draft standards should, of course, be tested and refined with concrete implementations using real education data in real data exchange scenarios.

Defining the Architecture

The architecture of the prototype proved to be highly successful in supporting both rapid and reliable development. Data is transformed from the source data in four stages, as depicted in the following figure:

1. Source data is transformed into XML as specified by the standard CDM XML schema.
2. Data in the XML standard is transformed into a relationally-modeled operational data store (ODS) directly modeled on the CDM reference model.
3. Data is transformed into a dimensionally-modeled data warehouse (DW).
4. The presentation layer retrieves data by calling a service interface to populate metadata that powers the dashboards and drilldowns.


Presentation Layer
Data Sources
ETL
XML
ETL
Operational Data Store (Relationally Structured)
ETL
Data Warehouse (Dimensionally Structured)
Service Interface
CDM Reference Model

Here's what went well

This architecture proved particularly effective at isolating the various concerns of the data transformation, as follows:

- The syntactic and semantic particulars of a data source were isolated in the mapping of the data source to the CDM XML standard. At this stage, the ETL (extract, transform and load) developer ensures that the input XML exactly matches the semantics of the CDM standard. As new sources of similar data are integrated, only this first stage needs to be developed; the remaining stages are unaffected.

- The ODS focuses on storing the LEA data, adhering to the CDM semantics but independent of any of its downstream uses. The ODS represents the *standardized* view of the raw LEA data in relational form.
- The DW is designed for efficient access for visualization and reporting.
- The presentation layer is optimally structured for the various dashboards and drilldowns, retrieving its data through a generalized service interface.

This separation of concerns served the team well time after time. For example, when the data format used by the testing vendor to send results to the ISDs changed, there was minimal impact "upstream" from the raw data, so the change was easily accommodated.

Here's what went not-so-well

Existing tools to support automated build and test of software are not tuned to support this style of multi-step data transformation. The team had to invent and implement custom infrastructure to develop and test the architecture.

Lessons to carry forward

The Release 1 architecture that separates the various transformation concerns was, so far, highly successful. Future releases will test the architecture's capability to scale to support multiple districts, to be expanded to support additional metrics, to incorporate different data sources, and to be extensible for new security features.

Development Process

Because the R1 prototype development required the rapid development of over 125 student and campus metrics, it was evident that a factory approach to the software development was required.   The development process was tuned to the four-stages-of-transformation architecture, integrating concepts of software factories, agile, and test-driven development.

The four-stage architecture clearly partitioned the work into four types of tasks:

1. Mapping and ETL development of raw source to XML per the CDM standard;
2. Mapping and ETL development of the XML to the ODS schema, where both are based on the CDM reference model; initially, this step also included creation of the ODS schema;
3. Mapping and ETL development of the ODS schema to the DW schema; initially, this step also included creation of the DW schema; and
4. Development of the user interface and associated presentation layer metadata.

Based on team's desire to incrementally develop the system driven by the metrics, the team selected a Kanban flavor to the software factory. *Kanban* is a manufacturing concept related to lean and just-in-time production. Kanban looks to streamline the manufacturing process by controlling the rate of production by the demand and limiting the works in process.[3]Hence, the process was to select a metric (or a cohesive set of metrics) and implement that metric end-to-end, from raw data to the dashboard user interface.

Because of the large number of metrics required, the team set themselves a goal to be capable of implementing a single new metric, end-to-end, in a single working day. The initial need for infrastructure and architectural scaffolding development plus the need to make refinements to the data model required more time for the first metrics. After the initial efforts were complete, the goal of 1-per-day metric development was generally achieved or exceeded.

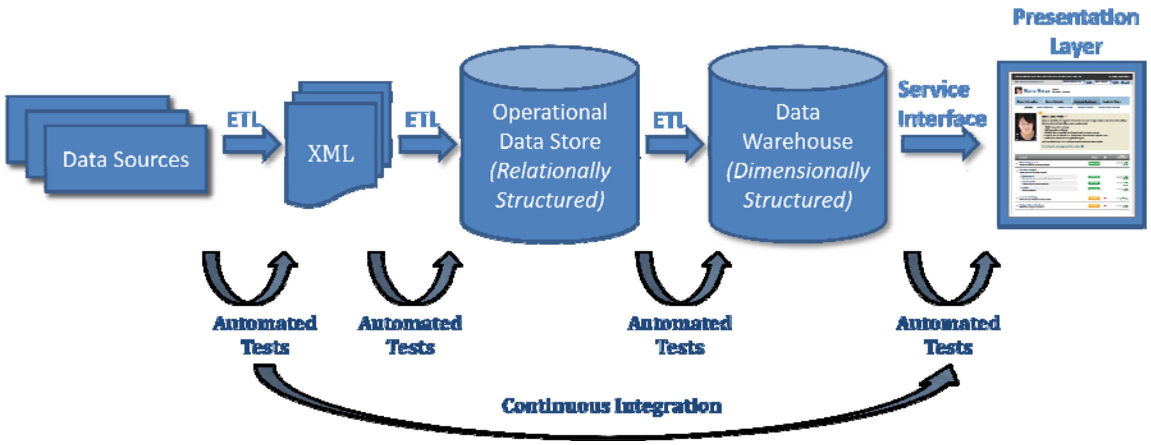
The exact definitions of the metrics were still very much in flux at the start of development – and definitions continue to be refined. The team implemented an agile development approach to accommodate the fluidity of the requirements. Agile software development refers to a category of software development methodologies employing an iterative process, where requirements and solutions evolve through collaboration between cross-functional teams—in this case, between the analysis and development teams.[4]

To accommodate a rapid rate of requirements change while implementing the large volume of metrics, the team selected a highly aggressive iteration length of one-week release cycles. Each Friday, the team members reported on the week's progress along with a demo of new metrics and functionality.

The development team members were experienced in test-driven development (TDD) and related lean and agile practices. TDD is a software development technique in which developers write an automated set of test cases that define the function prior to development of the code that implements the function.[5]Although TDD typically is not practiced in prototype developments, the team decided that the payoff in reduced rework as a result of changing requirements to a complex system would warrant the time and expense. TDD typically is not associated with ETL development; however, conceptually, it seemed to be a good match because it forces the ETL developer to identify every possible input situation prior to coding the ETL.

The development team implemented a custom infrastructure to support the automated tests for the four stages as well as a continuous integration process, as depicted in the following figure.

[3] Wikipedia Foundation,*Wikipedia, the free encyclopedia*, http://en.wikipedia.org/wiki/Kanban.
[4]Wikipedia Foundation, *Wikipedia, the free encyclopedia*, http://en.wikipedia.org/wiki/Agile_software_development.
[5]Wikipedia Foundation, *Wikipedia, the free encyclopedia*, http://en.wikipedia.org/wiki/Test-driven_development.



Development began on February 1, 2010, with the basic architecture and infrastructure in place by mid-February.

The team received the initial student data in early March and quickly realized that the end-to-end development was dependent on first mapping the raw source data to the XML schema, particularly given that the database schemas also were being designed for the first time.

Two weeks of data mapping and loading were necessary to support the operation of the metrics-driven factory; by the end of March, the team was executing a repeatable test-driven process to implement new metrics. The goal to implement a metric end-to-end in one working day was accomplished in the first week of April (though by then, some of the raw data required to drive new metrics had been previously mapped and loaded into the data store).

With an end-of-May deadline rapidly approaching, the previous investments in the architecture, development infrastructure and process paid off. The bulk of the student metrics were completed by mid-May; the campus metrics, which were often structured as roll-ups of the student metrics, were completed during the last two weeks of May.

Here's what went well

Generally, the development process was successful:

- The *principles* of Kanban for driving the factory by the software metrics (demand) worked well, despite the fact that the team's implementation did not strictly adhere to the process all points in the project.
- Agile processes allowed metric definitions to be refined throughout the process.
- Test-driven development turned out to be particularly useful for ETL development. TDD forces deep semantic thinking about the education data and the associated metrics dashboards before any development takes place. In the team's case, because different

developers were responsible for different components, TDD required the team as a whole to have a relatively exact understanding of the data before development began.

- The team developed a software tool that it called an “Anonymizer” to mask all identifying properties of the students in the system (e.g., by replacing all student and parent names with realistic-but-not-real names, replacing address information, etc.). The purpose was that the team needed to use real academic records (to prove the concepts for which the prototype was created), but also needed to show the resulting system to a wide audience (to validate that the work was, indeed, useful to its intended user base), all while maintaining the privacy required for student information. Although the creation of the tool was initially thought of as incidental to the project, the investment in time to create the tool was returned many times during the course of the prototype development.

Here’s what went no so well

Even though the metric definitions were grounded in prior practice and research, and then thoroughly analyzed as part of the specification process, the team still found relatively major flaws in the logic once the metric was implemented and visible on the screen and again when the metric was applied to real student data.

Lessons to carry forward

Some lessons learned will inform the process in the future:

- The Kanban board that the team used to control the development process was not as useful as expected. The hypothesized Kanban process assumed small increments of work being done simultaneously when, in reality, much of the process was done serially. For example, as the team awaited the data from Lubbock, the team decided to accelerate the user interface development. Once the Lubbock data arrived, the data mapping was accomplished next.

- The mapping and validation of raw data, while always on the critical path, needs to be accelerated and accomplished first.

- For this release, a tester—separate from the ETL developer—created the ETL test cases. In the future, the ETL developer will develop the test cases prior to developing the ETL. This will allow the ETL developer to benefit from the deeper semantic understanding that generating the test cases provides.

- Because the UI remained fluid for much of R1, the team decided not to invest in automated UI testing. However, this is planned for the second release.

Acquiring the Source Data

The project's database specialist spent two half-days with the Lubbock ISD staff in January 2010 to identify the required tables in the underlying database of their SIS. Using the CDM as a guide, they jointly identified about 120 tables out of the 2,700 possible tables in the ISD's SIS. The proper tables were relatively easy to identify despite the lack of a data dictionary or other database-level documentation.

Lubbock exported these 120 tables of the SIS and sent them to the development team via a secure transmission. Test results for the state standard assessment TAKS; the national college readiness assessments SAT and ACT; and benchmark data also were sent. The total data set was 3.24 Gigabytes, representing the student data for 48,562 K–12 students in Lubbock ISD schools. The estimated total time required of the Lubbock information technology (IT) staff, including responses to subsequent questions and follow-up data requests from the development team, was fewer than 80 total person-hours.

Access to the raw data was crucial. With the raw data, the project team could address issues relating to the deep data semantics of the SIS data, refining the CDM and/or modifying the ETL appropriately. This also reduced the burden on the Lubbock IT staff by avoiding repeated complex data requests and the need for Lubbock IT do the data mapping.

For example, in the initial processing of a student profile and demographic data, the team did not find the student's address. A quick question to Lubbock revealed that their SIS data indicated which parent a student lived with, and that the address was tied to the parent. The team quickly adapted its interchange schema to include the parents and their relationship to students. The ETL was modified to pull the student address from the parent with whom the student lived. With access to the original raw data, the team could modify its processing without requiring additional work from the Lubbock IT staff.

Here's what went well

A critical factor for success was acquiring and transforming the raw data obtained from the SIS database tables. The approach minimized the need for clarification questions back to Lubbock ISD that would have slowed the process. The process also greatly accelerated the team's understanding of the data because the raw data could be inspected.

Here's what went not-so-well

The impact of the delayed receipt of the Lubbock ISD data was underestimated. As discussed previously, the delay impacted the planned development process. This also resulted in a delay of having actual student performance data to inspect when questions of data or metric semantic surfaced.

Lessons to carry forward

During this process, the team uncovered several issues related to data availability and quality of data, as well as policies that made data transmission and metric development problematic. Many of these issues are not unique to Lubbock ISD, nor are they an indication that the Lubbock ISD staff is not monitoring data quality. Many issues are systemic, and are precisely the challenges that the TSDS system is designed to address. The specific examples below serve as thought-starters for other states and districts beginning a similar effort.

Some data proved to be elusive, as follows:

- Advanced Placement exam scores were not available in electronic form. They are delivered to Lubbock ISD in paper form only.
- For Career and Technical Education (CTE) students, only the CTE courses are tracked. The particular career objective — known as a cluster (e.g., automotive, transportation, hospitality) — is not tracked.
- Records of CTE students' professional certifications were not found in the SIS, nor were placeholders found for this data.
- Districts do not receive their accountability ratings for campuses electronically; the system would need to process the Texas accountability data from the Academic Excellence Indicator System (AEIS) to obtain these ratings.
- Districts do not receive data on GED completions for students who have left. GED completions factor into a campus' completion rate metric, which must be obtained from TEA's GED system.
- The SAT and ACT exams are administered throughout the year. Based on discussions with Lubbock, SAT scores are received at various times throughout the year, while the ACT files are received in December. Metrics relying on the SAT and ACT need to be adjusted for those students for whom the district has not received test scores. Furthermore, due to the receipt of files at different time periods, districts will need to upload this data in a timely manner to ensure the metrics are actionable.
- The TAKS electronic score files do not contain the metadata about the tests, objectives and assessment items. This data is available only in PDF form. Metadata mapping assessment items to student expectations is delivered in a separate submission after all TAKS testing is completed for the year.
- Free Application for Federal Student Aid (FAFSA) application data are provided to districts in paper form only. In Texas, the Texas Higher Education Coordinating Board (THECB) may be a better source to acquire this information electronically.

Issues discovered relating to data anomalies were as follows:

- Based discussions with Lubbock ISD, it seems to take about three weeks from the beginning of the school year to obtain stable enrollment figures.

- The TAKS and SAT columnar data formats change each year. Both test providers employ data practices requiring the re-definition of columnar layouts from year-to-year. For example, during this 2009–2010 school year, TAKS renamed the vertical score as the scale score and reported it in the columns for scale score, creating a point of possible confusion. It should be noted that the fixed column method of reporting is prone these type of changes; an XML-based interchange could avoid many of the problems.

- Many metrics are designed to evaluate trends over time. Comparisons of the spring 2010 raw district student enrollment data to the fall 2009 enrollment data that was officially reported found a difference at one campus of approximately 300 students (about 10 percent). The project team found that due to (a) the timing of the officially-reported (single snapshot in the fall), and (b) the fact that several campuses in Lubbock ISD have approximately 20 percent mobility within a school year, the enrollment data collected later in the year may differ significantly from the official accountability enrollment data collected earlier in the year.

- The metric that indicates whether a student has taken Algebra I had be changed to account for cases where a transcript did not show the student taking Algebra I, but did show that the student was taking a mathematics course later in the sequence (e.g., Algebra II, Pre-calculus, etc.). While this seems obvious in retrospect, the experience forced sharper thinking on the purpose of this and other metrics to properly deal with missing transcript information (e.g., for transfer students, for classes the student tested out of, for classes the student took in middle school, or other valid reasons for data to be “missing”).

Issues discovered that may relate to policies are as follows:

- Some students were found that had perfect daily attendance but high class-period absences. Lubbock ISD, like many districts, count children as present for the day if they are marked present in their homeroom or other designated class period. The fact that this phenomenon exists is a testament to the validity of a class period absence metric to identify all varieties of student absence issues.

- Other students were found to be marked absent for the day, yet were reflected as present in a few classes. The Lubbock ISD SIS records only marked absences (versus affirmatively recording both present and absent event). It is clear that some non-homeroom classes do not record attendance in the SIS every day.

- TAKS social studies has the highest percentage of test absences — twice the amount as other subjects across high school campuses. The project team hypothesizes that this is because passing social studies TAKS is not required for graduation.

- The student dashboard mockups include a picture of each child. While many stakeholders find the use of images compelling, many also expressed skepticism that schools or districts actually have image files available. To the team's surprise, Lubbock ISD has already piloted a system to house student image files — with little effort and at no additional cost to the district. Specifically: the district modified the agreement with its school photographer to require delivery of a standard picture format and file name based on the local student ID. A pilot was conducted with an elementary school and a middle school whereby a total of 830 student pictures were incorporated into the SIS. The team was able to extract and incorporate the photographs into the dashboard tool.

Mapping the Data

The source data was incrementally analyzed and mapped to the CDM XML standard. Separate concrete interchange schemas were composed from the library of types in the CDM-Core XML schema for various subsets of data (e.g., student data, enrollment data, attendance data, assessment scores, and student grades, etc.). This process was very effective because it allowed data to be moved into the DCD incrementally so that the metric development process could be driven incrementally. The structure of the CDM XML schema allows the easy creation of different sub-domain schemas to support different exchange scenarios.

Source data was mapped to the CDM XML standard using Altova MapForce. Mapping the SIS data and standardized test data took approximately 400 person-hours of effort -- including the time it took to analyze the source data tables and to create the initial DCD schemas. Mapping the data from its raw source to the XML standard and into the DCD had to be done prior to developing metrics from that data. Mapping the data from the raw source to the CDM XML standard provided an efficient way to load data from many different data sources while maintaining consistent meaning between sources.

The local district identification numbers were used for student identification in mapping the SIS; this process was highly reliable. Mapping of the TAKS and other tests used students' social security numbers.

For Release 1 of the prototype, the mapping activity was the first actual use of the CDM XML schema. In general, the mapping validated the structures and semantics in the CDM, making the mapping a straightforward yet time-consuming task. The live-data mapping process resulted in CDM re-factoring in (a) the complex set of entities and relationships associated with students, sections, courses, grades and credits, and (b) the three-tiered structures of test assessments and their metadata.

Throughout the process, care was taken when refactoring the CDM to keep the model consistent with its design goals of readability and source-system neutrality.

Questions relating to the semantics of the source data generally were answered easily via e-mail.

The CDM was designed for complete coverage of Texas's accountability data. Many accountability data elements are constructed from easily accessible operational data (e.g., attendance figures as reported for accountability are generated from operational attendance data). However, Texas's accountability system requires many elements that the team found difficult to deconstruct, usually because a single data element reflected multiple disparate concepts and/or had complex semantics. The CDM has a mechanism for "passing through" challenging elements through an extension scheme. Many of the difficult-to-deconstruct elements were found in the SIS data tables in the exact form required for accountability. This made the data mapping effort easy, but highlights that some accountability data elements are likely not constructed from data required to operate a district.

In addition, the data structures in the CDM mapped fairly well to the operational data at Lubbock ISD, but the values that go into those structures was occasionally district-specific and therefore required inquiry to derive the semantic meaning. Examples include:

- District-defined courses that could be substituted for state courses for graduation credit;
- Different campus-specific systems for section scheduling; and
- Extended discipline codes.

In all cases, the project team found generic solutions for dealing with these complex elements without compromising the source-system neutrality of the CDM.

Here's what went well

The data mapping activity provided to be less difficult than originally envisioned for several reasons:

- The investment in developing the CDM paid off during the data mapping phase. The CDM was the guide of what to look for in the raw data. It is much easier to look for something specific than to wander through tables trying to figure out what is important.

- The CDM XML schema supports the easy creation of interchange schemas. This allowed the data to be mapped and extracted incrementally. This allowed the team to "eat the elephant one bite at a time."

- There was a high level of cooperation from the Lubbock IT staff, who had significant technical and data expertise. Answers to questions were provided quickly and accurately.

Here's what went not-so-well

The raw data was transmitted about a month later than initially planned. As discussed previously, this meant the data mapping activity was on the critical path throughout Release 1.

Lessons to carry forward

The significant lessons for mapping raw student data are as follows:

- While it might be tempting to directly map the raw source data directly into a database, performing the transformation through the CDM XML standard forced critical semantic thinking that ultimately sped up the process.

- Incremental mapping and transformation are important when attacking a large education data set. The mechanisms in the CDM XML schema provide an efficient way to create interchange schemas.

Developing the Dashboard Metrics

The metrics went through four stages of development:

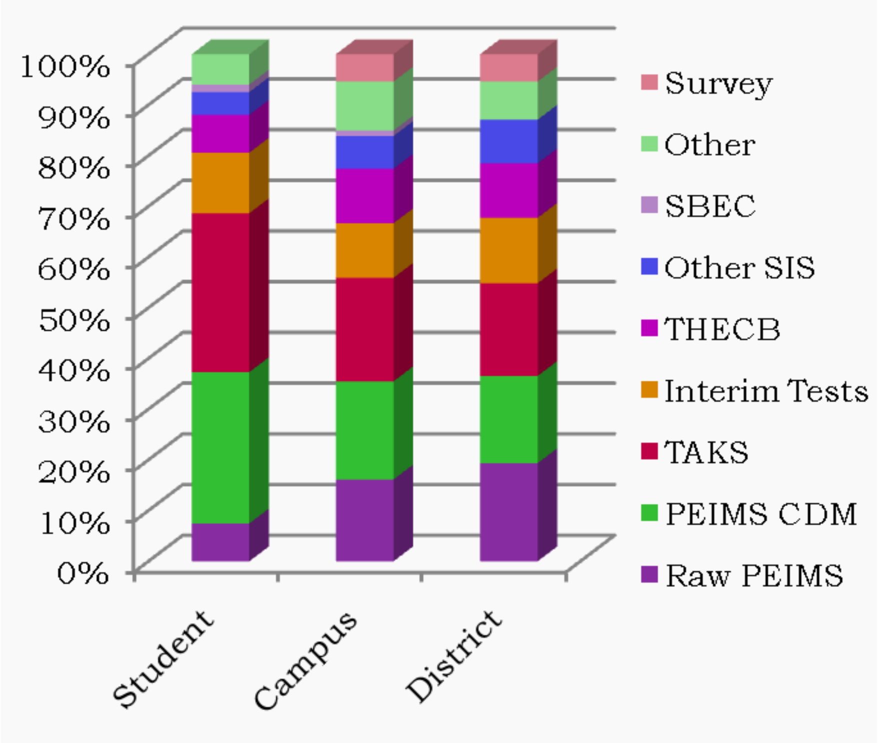
1. Metric scope and definition;
2. Metric detailed specification and mockup;
3. Metric refinement from a concrete implementation; and
4. Metric refinement from application to real data.

The metrics definitions were driven by a large amount of education research into leading and lagging indicators and by best practices reported by large districts with metric dashboard systems.

The TSDS metrics were defined during the first three months of the project, November–January. The team maintained the evolving metric definitions in a spreadsheet that included a brief definition of the metric plus other characteristics and categorizations. A few mockups portrayed the concept, but the examples used covered only a small percentage of the metrics. The concepts for the metrics were clear and relatively specific, but lacked formal specification.

Early in the initial definitional phase, the team worked through the issues of *data availability* (i.e., could data sources be identified that had the required data in a form that was extractable and would the districts provide access to that data) and *data periodicity* (i.e., how often data needed to be uploaded for individual metrics to be useful).

A core assumption was that some proposed student dashboard metrics could be derived from the exact accountability data that the districts were already submitting to the TEA. A gap analysis, graphed below, revealed that only a small percentage (10–15 percent) of the proposed metrics could be derived from the accountability data as submitted four times per year (identified in the graph below as “Raw PEIMS”).



However, when the more fine-grained source SIS data necessary to generate the accountability submissions (labeled "PEIMS CDM") was available, the number of suggested TSDS metrics fulfilled doubled to around 30 percent. That is, by using information in a finer grain which is also (generally speaking) more native to the information in the SIS, the system could produce more performance management metrics and also fulfill the needs of accountability reporting.

With the addition of TAKS standardized test scores, over half of the metrics could be derived.

The set of raw SIS data necessary to generate PEIMS, plus the TAKS scores became the defining characteristic for the metrics implemented during R1 and for the definition of the data needed from Lubbock. This definition was later amended to include historical SIS data. Ultimately, the R1 metrics were scoped to 145 metrics, determined by the raw SIS current data, plus historical data, plus TAKS, SAT and ACT.

In February and March, the analysis team created detailed *stories* for each metric. Each story included a mockup, a more complete specification of its computation and a mapping to its source data in the CDM. A typical metric story included three to five pages of write-up. Overall, the more detailed story write-ups led to:

- Greater specificity in definition, but tightly connected to the general information about a metric, like why the metric is important and what action is implied if the metric indicates a problem;
- More consistent user interface; and

- Elaboration of supporting items, such as metric trending details, thresholds and campus objectives.

The metric implementation activities in April and May led to the following:

- Further refinements in user interface;
- Refinements to metric titles and tag lines; and
- Expansion of some single metrics into multiple metrics (e.g., by subject, by grade, etc.).

The metric definitions were further refined in May and June based on reviews of the student and campus metric dashboards that were populated with actual data. The team flagged students who had seemingly inconsistent metrics. The team also identified certain campus metrics reflecting anomalies and/or inconsistencies. These refined metric definitions led to further refinements:

- Identification of new boundary cases (e.g., handling of students with late enrollment, students with incomplete transcripts, etc.);

- Discovery of new variants and situations (e.g., students that took the regular TAKS for some subjects and TAKS-M for others);

- Discovery of considerations that were previously missed (e.g., locally-defined courses that contribute to meeting graduation requirements); and

- Identification of situations with missing data (e.g., when a student took Algebra I in middle school and was currently taking Algebra II and the system did not have a full transcript, the system should infer that the student had indeed taken Algebra I).

The project team underestimated the amount of work that would be required to make the refinements at this stage. However, the inspection for data sensibility and quality proved to be critical. Until a system is stabilized, it generally cannot detect (and certainly cannot resolve) the vast majority of inconsistent or erroneous data— this can be accomplished only through manual inspection and is best accomplished by teachers, counselors and principals close to the students.

Here’s what went well

The team remained vigilant throughout the process for metric refinements, constantly searching for additional validations and expending significant resources to discover data anomalies that would ultimately result in improvements to the metric definition.

Here’s what went no so well

The number and degree of refinements to the metric definitions were greatly underestimated. What would seem to be complete, well-defined and well-founded metric at the start would require significant refinement, as discussed above.

Lessons to carry forward

Plan to refine metric definitions through application to a large sample of actual student performance data. Plan more time to inspect the results and validate each metric.

Designing and Developing the User Interface

The user interface evolved over the following steps:

1. An initial subset of notional dashboards illustrating the concept;
2. A set of detailed mockups for each metric; and
3. A user-interface implementation.

An initial set of conceptual dashboard designs (funded by the Michael & Susan Dell Foundation prior to the start of the technology work) established the overall look and feel of the dashboards. The foundation directed the development team to match this overall look and feel. The desire to align strictly to the design of the initial set of dashboards meant the development team had to custom program the user interface rather than use out-of-the-box business intelligence and dashboard tools. In drilldowns, which use standard line and bar charts, the direction to match mockups exactly precluded a simple use of charting libraries and required substantial effort to make the libraries work. The resulting dashboards have an extremely unique and clean look and feel in comparison with other education dashboard solutions that utilize out-of-the-box business intelligence tools.

Subsequent detailed mockups fixed a number of small inconsistencies that existed in the original mockups. Stakeholders identified user-interface issues, such as the unclear meaning of the summary indicators used on the student overview page. The first mockup looked to “compute” a summary indicator (i.e., derive a single red/green indicator from multiples). Stakeholders indicated that it was not obvious how these computations were being made. The next mockup displayed the number of red indicators in a red circle and the number of green indicators in a green circle. Stakeholders indicated that the meaning was not immediately obvious and that the overall visualization of the student performance was lost. The third mockup (that is reflected today) shows a better visualization reflecting each lower level indicator by a smaller indicator at the higher level.

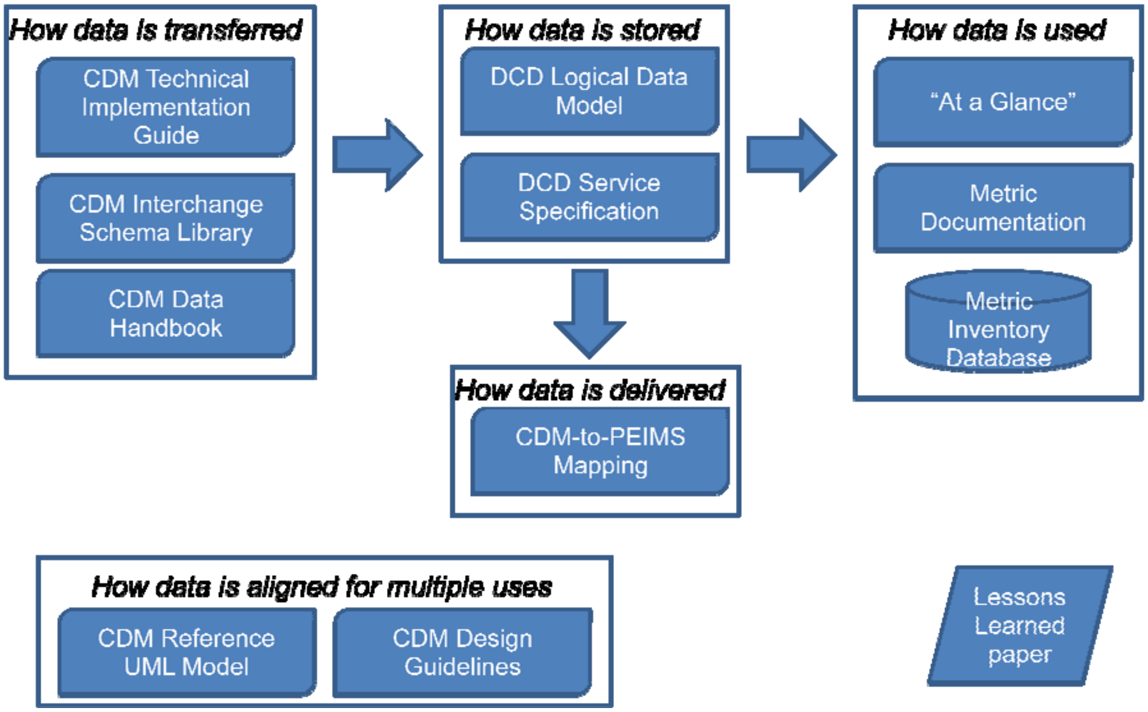
A number of detailed issues were resolved at the actual initial implementation time, including:

- Navigation;
- Search;
- Issues of screen clutter and readability;
- Handling of cases of missing data; and
- Some issues of color.

As discussed previously, the development team implemented a metadata layer to drive the user interface, which allowed development and testing of the user interface pages independent of the rest of the application. This approach supported the rapid development of a highly customized user interface. With such a high level of user interface development productivity, new visualizations were easy to implement directly from a notional sketch and to refine from there.

Documentation

Significant thought went into *architecting* the important documentation in manner that paralleled the architecture of the data system, as illustrated in the following figure:



The project team, working closely with the foundation, defined the desired documentation, including pointers to examples, to support the CDM and associated data standards. The only data-related challenge was that the detailed descriptions of the CDM and associated artifacts were too large to package practically in a document. For example, the generated documentation for the core CDM XML schema was more than 500 pages long. The project team concluded that it was better to deliver the detailed documentation electronically as Web pages. Any artifacts that remain as documents would be downloadable from the Web site.[6]

[6] The project documentation and other information artifacts from Release 1 can be found at http://www.districtconnections.com

The metrics documentation, however, proved to be more challenging to define. Many of the technical documents, such as the Reference Model or Technical Implementation Guide, had well-known templates or industry examples to follow. However the team could not find a template for the metrics documentation that served the project's needs. The challenge was to capture both the conceptual framework for a metric and an exact specification of data semantics and behavior.

The project team produced a set of metrics documentation used as specifications for the developers. The documentation was useful for developers and also appeared to be a viable basis upon which to build the final metrics documentation. The concept was to repurpose the documents to contain the description of the metric and how it was used in the same document as the technical and data details, with the goal of enabling mutual understanding between business and technical staff.

After several iterations to repurpose the documents, the research and analysis team worked to create exemplar write-ups for selected metrics. The exemplars included significant information about the basis for the metrics in terms of applicable research and best practices that were not part of the original technical documentation. As with the auto-generated schema documentation, the volume of metric documentation was too large for practical readability (greater than 500 pages), and was better suited for web-based delivery. Thus, the team developed a metrics database to hold the data and generate the electronic representation of the metrics information.

Here's what went well

Laying out the document architecture at the start of the project was highly successful. The CDM technical documentation had many good industry examples. As a result these documents were produced effectively and were of high quality from the beginning.

Here's what went not-so-well

Without good industry examples, the metric documentation wandered. As it was reviewed and edited by different member of the team over multiple iterations, inconsistencies and errors were introduced that had to be later resolved. The resulting metric documentation from this project may serve as a guide so that others implementing similar projects can use it as a discussion starter.

Lessons to carry forward

Spend time up front architecting the documentation and identifying the templates and examples that will serve as models. The forethought into the documentation architecture is often overlooked in projects. Even when the documentation is critically important, as it was in this project, documentation is often done as an afterthought.

Final Remarks

With a $10 million commitment from the Michael & Susan Dell Foundation, the development and implementation of the DCD — a primary component of the Texas Student Data System — is underway, and has so far successfully achieved its objectives.

Data analysis and technology experts from Double Line Partners and Cooper Consulting worked with Lubbock ISD to implement the first release of the DCD prototype, defining dashboards and associated metrics and developing the inner workings of the data system using raw data from Lubbock schools.

The first prototype release brought various challenges, which led to significant lessons learned. The observations discussed above will not only inform the DCD creation as the project moves forward, but also may serve to inform other states and districts embarking on similar efforts. Inevitable future lessons learned during the second and third phases of prototype release will allow further refinements to the DCD system and will further benefit public efforts to build off of proven best practices and the knowledge and experiences of others toward a national effort to develop successful data systems.

The DCD project team strives to support TEA's ultimate goal to update and enhance its statewide data system, continually improving the management of student performance and teacher instruction to benefit the students themselves and the surrounding community.

Glossary

ACT	Standardized test for college admission and placement (formerly American College Testing)
AEIS	Academic Excellence Indicator System. Texas reporting of education accountability data.
CDM	Canonical Data Model. A standard reference model for LEA education data.
CDS	Common Data Standards. An initiative to promote K-12 education data standards.
CTE	Career and Technical Education. A curriculum for vocational training.
DCD	District Connections Database. Operation data store and data warehouse for populating educator dashboards and generating compliance reports.
DW	Data warehouse. A data repository designed to facilitate reporting and analysis.
EDEN	Education Data Exchange Network. A set of education statistical reports gathered from state agencies by the US Department of Education.
ETL	Extract, Transform, Load. A process for loading a database or data warehouse.
FAFSA	Free Application for Federal Student Aid. Form to determine eligibility for student financial aid.
GED	General Education Development. A group of five subject tests which, when passed, certify high-school-level academic skills.
ISD	Independent School District. Managing entity for education separate from any government entity.
IT	Information Technology. The study, design, development, implementation,

	support or management of information systems.
LEA	Local Education Agency. A school district that operates primary and secondary schools.
LISD	Lubbock Independent School District.
MSDF	Michael and Susan Dell Foundation. The sponsor of this project.
NCES	National Center for Education Statistics. Collects, analyzes and make available education data.
NEDM	National Education Data Model. A conceptual representation of PK-12 education data.
ODS	Operational Data Store. A database designed to integrate data from multiple sources.
PDF	Portable Document Format. Open standard for document publishing and exchange.
PEIMS	Public Education Information Management System. The Texas repository for certified data for state and federal compliance and accountability reporting.
PESC	Post-secondary Electronic Standards Council. Promotes data standards for post-secondary education data.
SAT	Standardized test for college admissions (formerly Scholastic Aptitude Test).
SEA	State Education Agency. State-level government agency responsible for public education.
SIF	Schools Interoperability Framework. Transactional data sharing standard for interoperability between education information systems.
SIS	Student Information System. Software application for education establishments to manage student data.

SSIS	State-sponsored Student Information System. An opt-in, voluntary SIS offering for districts, hosted by the TEA.
TAKS	Texas Assessment of Knowledge and Skills. Standardized test for grades 3-11 in Texas.
TAKS-M	Texas Assessment of Knowledge and Skills - Modified. Standardized test for grades 3-11, modified for students with disabilities.
TDD	Test-Driven Development. Software development technique stressing test case development before coding.
TEA	Texas Education Agency. The SEA for the state of Texas.
THECB	Texas Higher Education Coordinating Board. Agency that oversees all public post-secondary education in Texas.
TPEIR	Texas P–16 Public Education Information Resource. The Texas statewide longitudinal data system for education data.
TSDS	Texas Student Data System. Next generation education data system planned for Texas.
UML	Unified Modeling Language. Standardized modeling language for software engineering.
XML	Extensible Markup Language. A textual data format for exchange standards.