xDValidator – Real-Time Data Validation and Standardization

1/23/2012

Data Quality and Standards

CPSI firmly believes that all systems should be based on standards, whether that standard is a data standard, a technology standard, or a platform standard. In K12, data standards are expressed as standard data definitions, code and value sets, business rules, and technical specifications. There are currently a variety of national standards, including CEDS, NCES, NEDM, PESC, EDFacts, IPEDS, EdFi, and SIF.

CPSI has the toolsets to ensure that all data standards are adhered to and that data quality checks (validations) are extensive and automated.

It is important for State Educational Agencies (SEAs) to adhere to established data standards to increase data interoperability, portability, and comparability across states, districts, and higher education. Many of the current applications and data sets at most SEAs do not meet any particular set of standards. Thus, the data sets need to be standardized when integrating the data. CPSI has the toolsets to standardize the data as it is extracted from these applications and data sets throughout the organization, and when integrating with other agencies.

Data standardization is very important, but data quality is just as critical. With typical file uploads, it is a very time consuming and intensive process. CPSI has the tools available to ensure that data standards are adhered to and that data quality checks (validations) are extensive and automated.

This document describes the approach that CPSI takes when developing the business rules related to data quality and defining the compliance to data standards. CPSI uses its xDValidator as the workflow engine and for defining and implementing business rules and data validations. CPSI uses its xDUA toolset for data transformation and for adherence to the standards. The xDStore also builds standard data schemas based on the industry standard data sets, such as CDS, NCES, NEDM, PESC, EDFacts, IPEDS, and SIF. The SEA can choose the data standards that are preferred, and the data standards can be mixed and matched between the Operational Data Stores (ODS) and the Reporting Data Warehouse, the Data Marts, and the Data Dashboards.

Data Quality

One of CPSI's core competencies is in the validation of data using the xDValidator toolset. All validations occur centrally at the SEA level and can be cloud-based as well. The CPSI Validation Rules Engine, xDValidator, performs real time validations at a rate of over 30,000 validations per minute. Validation at the application level is not required or necessary. The xDValidator Validation Rules Engine has an interface that allows the SEA to design and implement any validation on any tables and databases. The xDValidator is fully extendable through an API (DLL) or by implementing scripted C# code, SQL language, stored procedures and regular expressions. The system is designed to be extended and maintained by the end user. All errors and warnings are written to tables that can be accessed by any reporting system.

Once implemented, the xDValidator will be used to deploy the SEA's validation rules on both the xDStore ODS and the Reporting Data Warehouse. The validation rules provide four definable categories:

- *Actions*: Actions are generic code not specific to a table or a field. These actions can be written in C#, SQL, .NET Regular Expressions or as API calls into custom DLL's.

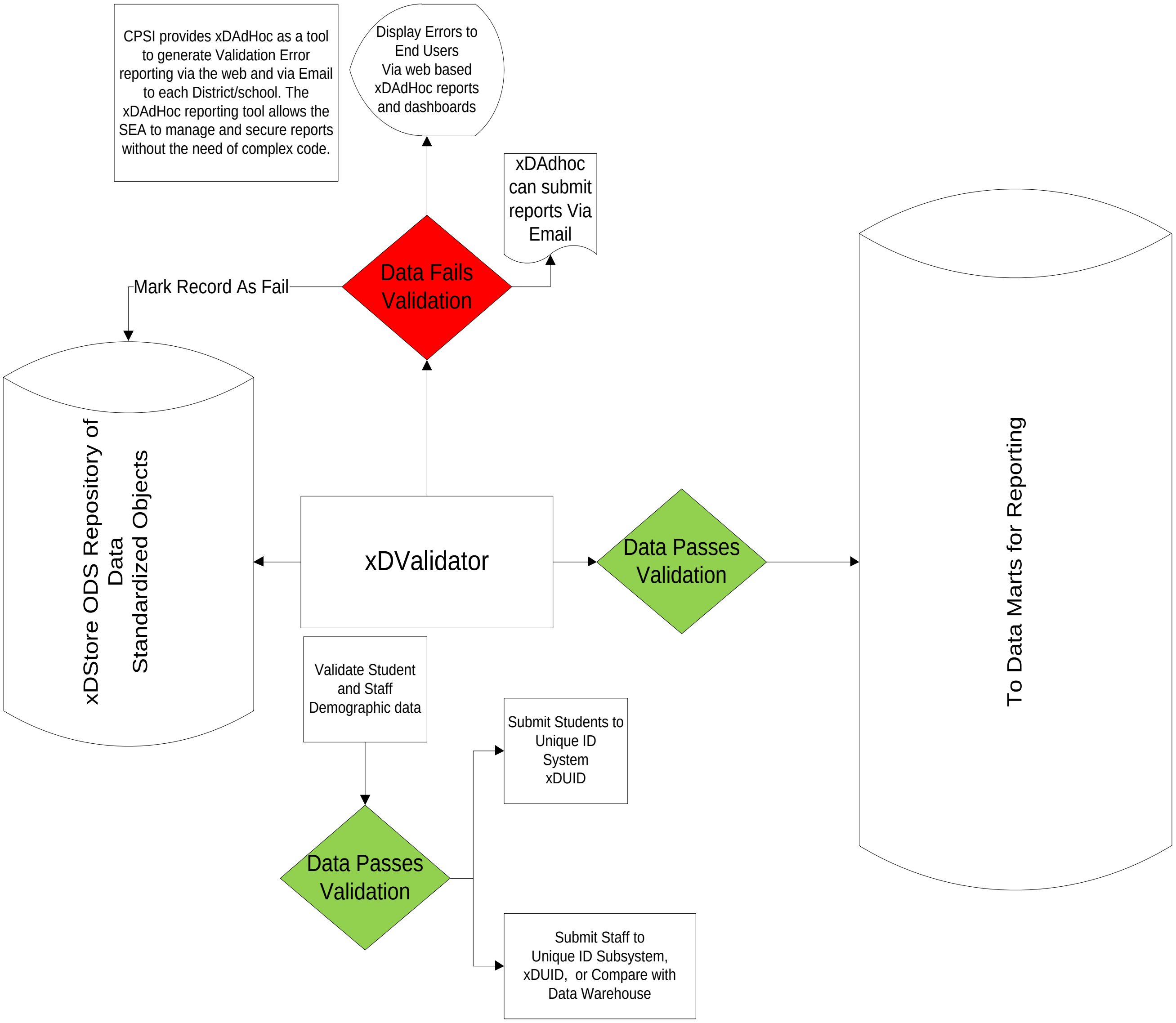
- *Items*: Items are Actions applied to particular fields in particular tables in the ODS.

- *Rules*: Rules provide the data selection criteria and the Workflow for the Items applied against the data.

- *Workflow*: Workflow is a collection of Rules managed as a Workflow and assigned to a thread of the Validation Service Engine on a specific server.

As data flows into the xDStore ODS, the ODS acts a large container that takes in all data, whether it is good or bad. The xDValidator is continually validating the data and flags the records that pass validation as good and marks bad records with errors or warnings. Records that fail validation are submitted to the error reporting process for action by the data owners.

The maximum data/traffic volume has high performance efficiency. The ODS agent can receive and process and validate 1000 to 2000 XML objects per minute per service instance depending on the message size. The average state deployment will have 5 to 10 Agent services running simultaneously. This will average 400,000 to 800,000 records per hour. The system is fully scalable, is dependent on the capability of the backend SQL server and is hardware dependent. In the Oklahoma deployment, we run five (5) services and we process about 400,000 records per hour on all 5 services under full load. The Validation Rules Engine can keep up with the ODS process regardless of the amount of data flowing through it.


CPSI provides xDAdHoc as a tool to generate Validation Error reporting via the web and via Email to each District/school. The xDAdHoc reporting tool allows the SEA to manage and secure reports without the need of complex code.
Display Errors to End Users Via web based xDAdHoc reports and dashboards
xDAdhoc can submit reports Via Email
Mark Record As Fail
Data Fails Validation
xDStore ODS Repository of Data Standardized Objects
xDValidator
Data Passes Validation
To Data Marts for Reporting
Validate Student and Staff Demographic data
Submit Students to Unique ID System xDUID
Data Passes Validation
Submit Staff to Unique ID Subsystem, xDUID, or Compare with Data Warehouse

Data Validation Process

1. Data is collected and stored in the xDStore ODS as a data repository of standardized objects and elements.
2. The xDValidator begins the data validation process.
    a. If the data fails validation:
        i. The record is marked as “Fail” in the xDStore.
        ii. Data errors can be sent back to the data owners via e-mail.
        iii. Data errors are displayed to the end users via a web based dashboard and reports created through xDAdHoc, which is provided as a tool to generate validation error reporting via the web and via email notification to each district.
        iv. The xDAdHoc reporting toolset allows the SEA to manage and secure reports without the need of complex code.
    b. If the data passes validation:

   i. Validated data is submitted to the current SEA Student Unique ID system or to CPSI's xDUID for assignment (process optional but provided if needed).
   ii. Validated data is submitted to the current SEA Staff Unique ID system or CPSI's xDUID for assignment or comparison (process optional but provided if needed).

3. Validated data can then be sent on to the Data Marts for reporting and analysis.

There are no limits on the validation reports that can be created. The validation system is integrated with Ad Hoc reporting capabilities. Each District/School/SEA Department will only see errors related to their data only. Users can view errors by object type, by error type, and by report impacted. Authorized users can create a variety of error reports that they desire without creating code. The system can be configured to allow the District/School/SEA Departments to create their own reports via the ad hoc capability. Reports can be scheduled, emailed, and exported as well as many other features.

Standards

As a company, CPSI decided to promote standards and best practices in its early days. Standards allow the educational organizations to more easily integrate, synchronize, and consolidate data from the various departments, exchange data with other departments or other organizations such as Higher Education, and to communicate effectively through shared report formats.

CPSI's toolsets enforce enterprise wide data standards, a common data vocabulary, and they maintain the use of standardized data in order to create and provide better reports.

CPSI defines the functions as follows:

1. Data Structure – the definition of data
   a. Define the databases that hold the data.
   b. Define the data that need to be gathered.
2. Data Architecture – the storage, movement, and retrieval of data
   a. Utilize the xDComposer to map the data from the source data silos and transport the data to the xDStore ODS.
   b. Utilize the xDStore as the ODS to maintain the data collection.
3. Master Data Management – the maintenance of consistent core data throughout an enterprise with all departments
   a. CPSI's toolsets provide for the mapping of data and data schemas to many different formats.
   b. The crosswalk of data from the published data to a recognized standard - such as SIF, PESC, EdFi, CEDS, and others - maintains that all data will be consistent.

4. Metadata – the management of data definitions and information about data
    a. The data is maintained with its metadata in the xDStore.
    b. The data is managed centrally.
5. Data Quality – the accuracy, completeness, and compliance of data
    a. Data is validated extensively in the xDStore using the xDValidator.
    b. Only validated data is allowed to pass through to the xDStore Data Marts and the xDVault LDS.
6. Data Security – the protection of data and the authorization to use it
    a. Data needs to be secure.
    b. CPSI's toolsets maintain encryption and other industry standard methods of security.
7. Data Reporting
    a. Data needs to be gathered into separate data marts for reporting purposes for the various departments. This will be done with the xDStore Data Mart tools.
    b. Data will be presented in a standard format in the reporting toolsets – xDTools and xDAdHoc.

All data that is sent to the SEA is staged in a series of secure folders or tables at the SEA. The xDComposer is the application that allows the mapping of the data within files or tables to the designated specification. The xDComposer supports any data standard currently used in education. It is fully programmable using .NET C# and no recompiling of the application is required. The xDComposer allows for code set transformations and advanced functionality during the mapping process. The xDComposer tracks and manages all REF ID's and cross links.

The xDComposer will convert the data that is sent via files to a standard also. Data is then sent on to the xDStore or to the xDZIS in a SIF environment. The xDComposer can play the role of a centralized SIF agent for all data applications/databases in the SIF environment. In all cases, the data is then routed to the xDStore.

The primary aim of this methodology is to standardize ALL data in the system.

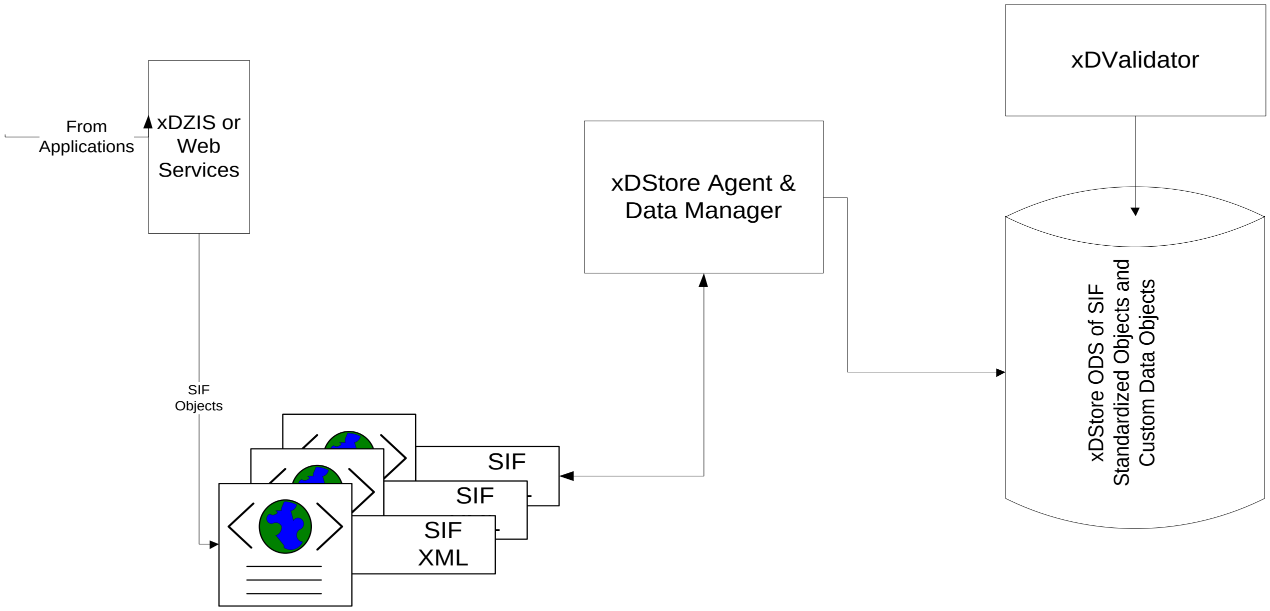
The PRIMARY aim of this methodology is to standardize all data that is collected to the specification of the data standard being implemented so it can be stored in an Object Based Staging Area, which is the xDStore Repository. The data is then moved to Data Marts and the LDS for advanced validation, processing and reporting.

CPSI uses the xDComposer to map the data from the local systems to the specified Data Model. CPSI's entire product set is capable of using any data standard by loading the XSD schema of the standard. For instance, if the SEA decides to standardize on SIF, then we will load and map the SIF Data Standard. The CPSI product set is capable of handling ALL data objects and elements through the SIF 2.4 standard. We can also add any Custom Object to the standard for mapping the data. Extended elements are fully supported as well. A gap analysis will determine what data is available, and what data the SEA will be collecting. The difference in the data will determine the data gap, and that will determine the schema of the xDStore and the Data Marts. Elements that are not easily mapped to the particular specification can be formed as Custom Data Objects for the purpose of data mapping.

The functional processes that will guide the services for the Data Mart utilizes the following method of assigning responsibility so that the proper groups have ownership over the process:

Define Data Governance Process	CPSI and the SEA will designate the “owners” of the data that will be published from each data silo.
	Through the initial meetings, CPSI and the SEA will define the participation of the departments, including the IT staff, administration, and others.
	CPSI and the SEA will determine the data elements to be used from each data silo.
	CPSI and the SEA will define the procedures to approve the data elements.
Implement Data Governance Process	The SEA will designate the data stewards, or those that will approve the data elements and mappings.
	The SEA will follow the procedures to define and approve the data elements.
Create, Collect, and Maintain the Data Standard	CPSI will utilize the agreed-upon data standard to standardize the data elements, attributes, and data schemas.
	CPSI and the SEA will add the required data by defining custom data objects and elements and adding them to the standard.
	CPSI will maintain the naming standards, data classifications, business rules, data models, data dictionary, and data format standards in the metadata tool. The SEA will continue to maintain the data over time.
Develop and Implement the Enterprise Metadata Architecture	The xDStore will be created using the data that has been collected from the data silos.
	The xDStore will maintain the metadata.
	CPSI and the SEA will create the first set of business rules in order to validate the data coming into the xDStore with the xDValidator.
Create and Maintain Master Data Management Standards	CPSI and the SEA will build the various data marts for the departments based on their data needs.
	Business rules will be maintained in future iterations by the SEA using the xDValidator.
Standard Data Reporting	The data marts will be maintained for reporting purposes, including the Data Marts and Dashboard Data Stores. EDFacts data marts can also be created.
	xDTools and xDAdHoc can be used as a common reporting toolsets for users throughout the organization. The SEA can also use any current reporting applications on the data.

State Side
xDStudio


xDValidator
From Applications
xDZIS or Web Services
xDStore Agent & Data Manager
SIF Objects
SIF
SIF
SIF XML
xDStore ODS of SIF Standardized Objects and Custom Data Objects

Data Standardization Process

Conclusion

An important part of K-12 education is the ability to collect and analyze data from a variety of resources, including at the SEA level. An essential element of longitudinal analysis and vertical reporting is the use of qualified and validated data. CPSI's technology helps meet the challenge of integrating data from districts, their schools, and the state into a meaningful timely and consistent manner.

The combination of the xDComposer, xDValidator, and xDStore as the application backbone, the robust configuration abilities, and a fully functional standardized data model provides the ideal platform to implement a data quality system and support longitudinal analysis and vertical reporting now and in the future.

Technology trends and policy initiatives in K-12 education have helped fashion a new generation of data-driven stakeholders that operate in real-time. Students, educators, administrators, policy makers and parents, have heightened expectations for the ways in which student progress is tracked and impact is measured within our school systems. The xDStudio technologies play an important role in enabling delivery on the promise of NCLB and meeting the 21st century data-driven challenges.