SLI Bulk Data Transfer – RFP Guidance

SLC Project Document
February 2nd, 2012

Change Log

Date	Version	Name	Change Summary
1/30/12	V1	EFloyd	Initial draft submission to SLC
1/31/12	V1.1	EFloyd	Incorporated A&M feedback, synchronized with 1/30 update
2/1/12	V1.2	Alok	Updated table 6.1 to refer to Data Ingestion Specifications
2/2/12	V1.3	SKrongard	Removed @interchange from parameter table
2/2/12	V1.4	EFloyd	Updated section 3.5 to correct the example CSV control file format

Table of Contents

1. Introduction

This document is part of a series of documents that contain specifications for application software and system procurement where integration with SLI technologies is required. This *Draft Specification Document* provides a draft view of a future SLC released document and is intended to be referenced in vendor RFPs. As of this writing, the SLI standards are still in development. The technical information in this document should be considered preliminary.

This document provides specifications for Bulk Data Transfer.

1.1. Structure of Document

The draft specification is divided into five sections:

- **Overview** – Provides a broad description of the SLI technology upon which the requirements are based, including use case summaries.
- **Integration Approach** – Describes one or more approaches for integrating with a core SLI technology.
- **Relationship to Other Standards and Technologies** – Describes how a proposed solution is expected to make use of, or facilitate the use of, other SLI technologies.
- **Configuration Options** – Discusses areas of potential configurability.
- **Standards and Technologies** –Identifies applicable standards and technologies' and specifies their applicability to this standard. This section also identifies related projects, initiatives, and organizations.
- **Constraints** – Specifies constraints and exclusions that a proposed solution must satisfy.

2. Overview

This section provides an overview of Bulk Data Transfer.

2.1. What is Bulk Data Transfer?

Bulk Data Transfer is a general term that refers to the process by which batches of information are uploaded into the SLI Data Store or downloaded for local use. The upload process is also referred to as, "Data Ingestion." The primary focus of this document is on the Data Ingestion aspect of Bulk Data Transfer.

2.2. The Objective

The objective of the Data Ingestion process is to provide a way to capture locally-maintained education data and load it into the SLI Data Store while assuring that the integrity of the data in the Data Store is not compromised.

2.3. Use Case Summaries

Selected use case summaries are provided below in order to facilitate a general understanding of Bulk Data Transfer.

Title	Summary
Mercury County School District (MCSD)	MCSD is a small district whose current SIS vendor has announced end-of-life for the product version that the district currently uses. MCSD has selected a hosted solution that is based on the SLI Data Store. The current SIS product is widely deployed and data ingestion facilities for it are available free of charge from the SLC shared resources center. MCSD joins the state Realm, procures a high-end desktop PC to use for data staging, installs and integrates the SIS-specific data ingestion tools, and begins nightly data synchronization between the local SIS and the SLI Data Store. The district runs the old and new systems in parallel for three months while verifying reports from the new system and customizing the batch data feeds to other local MCSD systems.
Neptune Consolidated Schools (NCS)	NCS is a large urban district with a skilled, dedicated IT staff in-house. The District CIO has been following the progress of the SLI and has begun to see opportunities to leverage free or low cost SLI-based applications to provide services that would be prohibitively expensive under the expected budget constraints in coming years. These opportunities depend on the availability of NCS data in the SLI Data Store. The CIO initiates a project to work with the SLC and NCS's SIS and H/R application providers to implement data extraction and SLI data ingestion processes.

Title	Summary
Ort Cloud Consortium	A large metropolitan area is partitioned into 20 school districts of varying sizes. Districts are fiercely independent and wary of SEA control. The SEA has recently standardized on SLI as its future supported platform for all state statistical reporting and has allocated resources to develop and maintain SLI-based reports. Twelve of the smaller districts use the same SIS vendor. This vendor, recognizing an opportunity to reduce its support costs, allocates resources to implement data extracts suitable for SLI ingestion, and argues convincingly for district authorization to provide data to the SLI. Central to the argument is that each district has total control over the availability of its data, and that this control is enforced by an independent agency, the SLC, not by the SEA. Ten districts agree to cooperate, join the state Realm, and begin transferring data to the SLI Data Store.

3. Bulk Data Transfer Approach

The Data Ingestion approach depends on the type of product or service being proposed.

Product / Service	Potential Integration Role
Student Information System (SIS)	Map internal data elements to SLI/Ed-Fi data entities. Provide data extracts that conform to one of the accepted Data Ingestion formats.
Human Resources / Payroll System	Map internal data elements to SLI/Ed-Fi data entities. Provide linkage to the local IdP directory to facilitate Teacher and Staff sign-on. Provide data extracts for Staff and Teacher data that conform to one of the accepted Data Ingestion formats.
System Integration / Consulting	Recommend products that support SLI Bulk Data Transfer. Educate State (SEA) and Local (LEA) Education Agency technical staff in SLI facilities and processes. Assist an SEA or group of LEAs to establish a Realm. Act as Interim Realm Administrator. Assist LEAs in installing and configuring SLI Data Ingestion software and processes. Assist SEAs, LEAs, and vendors in developing data mappings and data extract software.
Software Development	Build SLI Bulk Data Interchange capability into new and existing products. Develop custom data extract software for Data Ingestion.

In addition to the process and Landing Zones for importing large amounts of data, there will be a process to allow SEAs and LEAs to export large amounts of data from the SLI data store. All data that is controlled by the SEA or LEA, both core and custom, will be exportable. Like the Data Ingestion process, the Bulk Export requests will be handled asynchronously. Data should be processed and available in the SLI within hours of the data load being initiated (exact SLAs to be determined).

The following sections provide further details of the Data Ingestion process. Additional information on SLI Bulk Data Transfer is available at http://www.slcedu.org.

3.1. Realm Onboarding

The process of enabling districts for Bulk Data Ingestion is called “Onboarding”. For the purpose of loading information, districts within a state are organized into one or more[1] “Realms”. A Realm is a hierarchical segment of a state school system comprised of one or more Districts, entities belonging to those Districts (Schools, Teachers, Students,

[1] For SLI Version 1, all districts within a state that upload data to the SLI will be members of a unique realm created for the state.

Courses, Assessments, etc.), and their associations. A Realm is the smallest unit of onboarding.

A Realm is created when a group of District SLI Super Administrators contract together to provide certain defining information to the central SLI administration repository. For each Realm, the following information needs to be collected and stored in the central SLI administration repository at the time the Realm is created.

- Realm administrator name, institution and contact information.
- A list of Super Administrators for Districts within the Realm to be added to SLI Platform Directory.
- Default Realm Identity Provider.
- Information necessary to generate a unique Realm ID.
- Version of WGEN's extended Ed-Fi Schema to be used.
- Expected number of records to be loaded for each entity during initial ingestion.
- Expected yearly growth of records for key entities, e.g. Student, StudentAssessment, Session.

The Realm Administrator can then designate and provision IT Administrators for each district. Once IT Administrators for districts within the Realm are identified and provisioned, they are securely notified of their login credentials, the SLI-hosted directory to be used for authentication, the unique Realm ID, and a "landing zone" URI for data files.

3.2. Ingestion Process

SLI will ultimately support three different ingestion vectors:

1. A SFTP interface to upload ingestion jobs to a designated landing zone.
2. A web-based submission interface.
3. A SIF agent interface.

All ingestion vectors provide a way to authenticate, create and submit a job, monitor its progress throughout the ingestion pipeline, and to be notified of any errors encountered along the way. The specifics of ingestion jobs initiated through the web interface and an event-based ingestion via a SIF agent are currently TBD.

Batch upload proceeds as follows:

1. An SLI Administrator prepares a batch job consisting of a job control file and a collection of data files exported from source systems.
    a. The data files may be presented in either of the two supported formats: XML data that adheres to one of the allowed Ed-Fi interchange schemas, or appropriately formatted comma-separated values (CSV).
    b. The job control file contains the ingestion configuration parameters and a listing of data files associated with the ingestion job.

2. An IT administrator, or a contracted data service, packages an ingestion job as an encrypted .zip file, signs on to the SLI Ingestion SFTP service, and transfers the file to the Realm's designated ingestion landing zone.
    a. Once transferred, the job will be assigned a job number, ingestion processing will begin, and a job progress log file will appear inside the landing zone.
    b. While an ingestion job is in progress, the IT Administrator will download log files to monitor job status.
3. Following a full pass of ingestion, the job log file will contain a set of warnings and errors generated in the process, as well as the information about the number and types of records ingested.

SLI will provide users with the appropriate permissions the ability to audit ingestion history. A log of ingestion jobs, and the users who initiated them, will be kept for future reference. This functionality is currently under definition with additional details TBD.

3.3. Control File Format

The control file contains a row of comma-separated values for each inbound data file that allows for the basic integrity checking of these files.

The row format is currently defined as:

<file format>,<interchange>,<version>,<entity>,<file name>,<file checksum>

where:

<file format>	Specifies the file format. At the time of this writing, ***csv*** and ***edfi-xml*** are the only supported file formats.
<interchange>	Applicable Ed-Fi interchange schema name.
<version>	Interchange schema version identifier.
<entity>	For CSV, represents the type of entities(s) contained in the file. For XML, the value should be left blank. This field is case-sensitive.
<file name>	Specifies the file's name. File names are case insensitive and should not contain any OS-specific path delimiters.
<file checksum>	File's MD5 checksum expressed as 32 hexadecimal digits.

The control file format allows for specification of job-level parameters. These are specified in the control file as line entries preceded with the "@" symbol. Job parameters appearing in control files are parsed using the specification for Java Properties after the leading "@" is stripped. Parameters that do not require a value are treated like flags.

Command	Effect
@dry-run	Indicates that the results of ingestion processing should not be written to the core data store.

The full set of control file parameters is currently TBD.

A sample job control file may look as follows:

@dry-run

edfi-xml,StudentEnrollment,1.0,,data.xml,756a5e96e330082424b83902908b070a

3.4. XML Data Format

Ed-Fi Interchange Schemas define XML representations of particular data spaces, or groups of entities and associations, for transport between systems. In most cases, different interchange schemas will be used to reflect different use cases, such as different groups of source systems. Ed-Fi defines a Core Schema that provides a library of building blocks, which are referenced from the interchange schemas. Ed-Fi defines thirteen standard interchange schemas that are be used by SLI. They are:

- AcademicMetadata
- EducationOrganization
- EducationOrgCalendar
- MasterSchedule
- StaffAssociation
- StudentAssessment
- StudentAttendance
- StudentCohort
- StudentDiscipline
- StudentEnrollment
- StudentGrade
- StudentParent
- StudentProgram

Each XML file, which is part of a well-formed ingestion job, is validated against an Ed-Fi interchange schema.

3.5. CSV Data Format

XML interchange schemas defined by Ed-Fi allow ingestion of a single XML file which contains entities of multiple domain types. The same approach is not possible with CSV;

CSV files require the same tabular format for all rows, which means that all entities in a single CSV file must be of the same type.

Thus a CSV ingestion job will contain multiple CSV files, each containing a set of entities covered by the same interchange schema. For instance, if a CSV ingestion job contains Student, Parent, and StudentParentAssociation entities from the StudentParent interchange, then it will contain a separate CSV file for each one of the ingested domain types. A control file for such an ingestion job may look as follows:

csv,StudentParent,1.0,Student,students.csv,95b3b66973da25541e7939753b1abf04

csv,StudentParent,1.0,Parent,parents.csv,95b3b66973da255b1abf0441e7939753

csv,StudentParent,1.0,StudentParentAssociation,spassoc.csv,953da25541e793975b3b66973b1abf04

A single ingestion job cannot contain a combination of CSV and XML data files.

3.6. Dependencies

The following interchange dependencies must be honored for ingestion:

Interchange Name	Direct Dependencies
EducationOrganization	
MasterSchedule	EducationOrganization
AssessmentMetadata	MasterSchedule
EducationOrgCalendar	EducationOrganization
StaffAssociation	EducationOrganization, MasterSchedule
StudentParent	
StudentAssessment	AssessmentMetadata, StudentParent
StudentAttendance	EducationOrganization, StudentParent
StudentProgram	EducationOrganization, StudentParent
StudentCohort	StaffAssociation, StudentParent, StudentProgram
StudentDiscipline	EducationOrganization, StudentParent

Interchange Name	Direct Dependencies
StudentEnrollment	EducationOrganization, StudentParent
StudentGrade	EducationOrgCalendar, StudentParent, StudentAssessment

The EducationOrganization should be the first interchange loaded, as it contains the following key entities on which most other interchanges are dependent:

- State and Local Education Agencies (StateEducationAgency and LocalEducationAgency entities)
- Schools
- Course information (Course, CourseOffering, and Location entities, etc.)
- Programs

The MasterSchedule interchange, which depends on entities loaded as part of EducationOrganization, should be loaded next. Once the EducationOrganization interchange has been successfully loaded, the following interchanges can be loaded in any order:

- EducationOrgCalendar, which contains calendar information necessary for loading grade-related entities.
- StaffAssociation, which contains Staff, Teacher, and related association entities.
- StudentParent, which contains Student and Parent entities and on which all remaining student-related interchanges are dependent.

After the three key interchanges above have been uploaded, the following interchanges can be processed:

- StudentAttendance
- StudentCohort
- StudentDiscipline
- StudentEnrollment
- StudentProgram

The loading of the remaining assessment and grade-related interchanges must be performed in the following sequence:

1. AssessmentMetadata
2. StudentAssessment
3. StudentGrade

3.7. Custom Data

While the exact design is not yet finalized, SLI will also provide the ability for applications to store custom data that is not defined by the SLI Core Entity Model. This data is considered "opaque" to SLI, in that SLI does not need to and cannot know anything about the structure of the custom data. SLI provides for simple storage and retrieval of that data by applications in order to facilitate application development without requiring the usage of a local application data store.

SLI will require that custom data be related to, or "linked", to existing entities in the SLI CEM. Access to custom data will then inherit the security permissions of the entity to which the custom data is related. For example, custom data that is related to a particular Student record will only be visible to users that are authorized to view that existing Student record. Custom data should not contain personally identifiable information (PII).

Custom data, by default, will be kept private to the application that created that data. However, we expect to also provide a "shared" custom data space, which allows for SEAs and/or LEAs to permit certain custom data to be shared across applications, but authorization to use that data will still be determined explicitly by the SEA or LEA for each application. An example of use case for this functionality is an SEA who has multiple custom applications, built on SLI that need to access Assessment and/or Student metadata that is not captured in the SLI CEM.

To enable applications to make use of shared custom data, SLI will also enable the storage of metadata that describes the structure of that custom data. This allows, for example, an XML Schema (xsd) to be stored to describe custom data so that other applications may then parse that data and use it for processing or presentation.

4. Relationship to Other Standards and Technologies

Data Ingestion makes use of, and facilitates the use of, the following standards and technologies.

4.1. SLI Application Programming Interface (API)

The Application Programming Interface provides the mechanism to transport information between the SLI and consuming applications. Interactive application access to the SLI Data Store is strictly governed by this API. The overall objective of the API is to provide a stable, well-defined interface for software developers. The API is a real-time transactional interface intended for interactive applications, while Bulk Data Transfer is intended for periodic batch data interchange.

The SLI API is the subject of another Draft Specification document.

4.2. SLI Identity Integration

For Batch Data Transfer, operators are required to authenticate against the SLI Platform Directory.

A Realm Administrator is a special SLI operator role which is authorized to perform administrative tasks for the Realm. As part of the Onboarding process, the Realm Administrator's and Super Administrators' accounts are created in the SLI Platform

Directory, data store and ingestion resources are provisioned, and the Districts' federated directory services are configured to communicate with the SLI Identity Service.

A Super Administrator has the full set of IT Administrator permissions by default, but can also delegate authority to ingest District's data to other IT Administrator(s) in that District. The Realm Administrator can provision these IT Administrator accounts within the SLI Platform Directory when granted this permission by a Super Administrator.

The SFTP server authenticates the Realm Administrator, Super Administrators, and IT Administrators against the SLI Platform directory, so Super Administrators and IT Administrators must login with the credentials provided to them at the time of Realm provisioning. The user submitting ingestion jobs must be associated with the landing zone URI and be provisioned previously by the Realm Administrator to have IT Administrator role/permissions for the Realm. Subsequently, the Realm Administrator can prevent said user from submitting ingestion jobs by removing IT Administrator role/permissions from the corresponding IDP's records.

The exact mechanism to tie SFTP username and password and the user identity stored in the SLI Platform IDP and the user entities stored in the SLI CEM are currently under definition, with additional details TBD.

Identity Integration and Management is the subject of another Draft Specification document.

4.3. Core Entity Model

Data prepared for SLI Batch Data Transfer is organized according to the SLI Core Entity Model (SLI CEM). The SLI CEM is an abstract, technology-agnostic representation of the K-12 education information domain. The model includes entities that are easily recognized: school, student, course, section, among others. Those entities contain attributes that are also easily recognized, though a complete listing of entities and attributes are beyond the scope of this document.

SLI CEM contains entities along with the relationships that define how the entities interact with one another. Each entity includes a sufficient number of attributes to make the model applicable to real-world data. SLI CEM focuses on granular information rather than aggregate statistics. In addition, the model includes information that is necessary to produce aggregate and other types of statistics.

SLI CEM is expressly focused on representing the instructionally relevant classroom-level student and educator-focused data that educators can use to differentiate instruction, support individual student, need and help to improve student outcomes. The SLI shares this priority use case with the Ed-Fi initiative. For this reason, the SLI CEM is based on the Ed-Fi Logical Data Model.

5. Configuration

Areas of potential configurability include:

Area	Potential Configuration Items

Area	Potential Configuration Items
Onboarding	Realm Administrator, district Super Administrators, identity provider, schema version, IT Administrators
Transfer Management Tools	credentials, landing zone URI, data sources, field mappings, file format, schema version, local staging directories

6. Standards and Technologies

The following standards and technologies are applicable to this specification:

Standard / Technology	Applicability
Ed-Fi Design Guidelines. Version 1.0. http://www.ed-fi.org/wp-content/uploads/2011/06/Public-Ed-Fi-Design-Guidelines-1.0-111111.pdf	Introduction to the Ed-Fi unifying data model (UDM) – notation, rules, conventions, domains, core schema organization, extensions.
Ed-Fi – Unifying Data Model. Version 1.0. http://www.ed-fi.org/wp-content/uploads/2011/06/Public-Ed-Fi-Unifying-Data-Model-1.0-111111.pdf	The Ed-Fi UDM, organized by domain. For each domain, the graphical UML model is provided as well the definitions for the entities and associations.
Ed-Fi Interchange Schemas http://www.ed-fi.org/wp-content/uploads/2011/12/Ed-Fi-Sample-Interchange-Schemas-with-Data-1.0.zip	Ed-Fi Interchange XML schemas.
Ed-Fi Core Schema http://www.ed-fi.org/wp-content/uploads/2011/12/Ed-Fi-Core.zip	Ed-Fi-Core.xsd
Java Properties Specification. http://docs.oracle.com/javase/6/docs/api/java/util/Properties.html#load%28java.io.Reader%29	Control file job-level parameters are parsed according to this specification.
Secure Shell (SSH) http://en.wikipedia.org/wiki/Secure_Shell	Underlying secure protocol upon which SFTP is based
SSH File Transfer Protocol http://en.wikipedia.org/wiki/SSH_File_Transfer_Protocol	Protocol used to transfer data securely to and from an SLI Landing Zone
PUTTY, Tunnelier http://www.putty.org/	Free SSH / SFTP client software. Tunnelier, in particular, provides a convenient “explorer” style SFTP user interface.

6.1. Related and Affiliated Efforts

Initiative / Project / Organization	Applicability
SLI Application Programming Interface (API) http://www.slcedu.org	Fully-integrated SLI Portal applications use the API to interact with the SLI Data Store

Initiative / Project / Organization	Applicability
SLI Core Entity Model http://www.ed-fi.org/wp-content/uploads/2011/06/Public-Ed-Fi-Unifying-Data-Model-1.0-111111.pdf	The SLI Data Store is organized according to the Core Entity Model.
SLI Data Ingestion http://www.slcedu.org	Bulk Data Ingestion and Validation are the means by which SEAs and LEAs store information in the SLI Data Store. The related file is named SLI Data Ingestion Specifications.
Identity Integration and Management http://www.slcedu.org	Ensures security and access control
Learning Resource Metadata Initiative http://www.lrmi.net/	Tagging standards to facilitate content management and discovery
Learning Maps	Pathways through learning objective standards

7. Constraints

To be compliant with this specification, solutions will be subject to the following constraints:

1. Ingestion order must honor interchange dependencies and normative constraints such that referential integrity is preserved.
2. Staff,Teacher, Student, and Parent entities ingested as part of the StaffAssociation and StudentParent interchanges must be associated with user sign-on credentials in the local identification provider directory. The exact mechanism to make this association is TBD.
3. The project will publish a set of data modeling guidelines that support SLI-aligned definitions and constraints.