Improving Data Quality with the eScholar Complete Data Warehouse

Although many data warehouses are initially constructed to meet state and federal reporting requirements, the educational agencies with a data warehouse can more quickly and effectively select, gather, analyze, manage, and improve their data to establish credibility, improve performance, and reduce costs. As stated in the Malcolm Baldrige Education Criteria for Performance Excellence, critical components of how organizational information and knowledge are managed require that data:¹

- Accuracy
- Integrity and reliability
- Timeliness
- Security and confidentiality
- Currency with educational service needs

The Challenge

Based on eScholar’s implementations in over 3,500 school districts across the country, common problems in collecting and using data include lack of:

- Clearly articulated goals, strategies and data used to measure success
- Common definitions for important data elements
- Time, training, and “adhered to” policies to collect and verify data
- Required staff and other resources necessary to support the data warehouse
  Support from key stakeholders, including policy-makers, administrators, teachers and parent

The Solution

State and districts need a plan to strategically use their data and a data warehouse that provides:

- A standardized, automated process incorporating best practices
- Continuous alignment with national standards
- Fast implementation and performance
- Accurate data
- Low maintenance costs

eScholar has developed the RADAAR approach shown in the table on the follow pages that encapsulates best practices in data cleansing.

¹ The Malcolm Baldrige 2006 Education Criteria for Performance Excellence Category #4 Measurement
<table>
<thead>
<tr>
<th>RADAAR Category</th>
<th>Sub-Category</th>
<th>Best Practice</th>
<th>Increase Data Accuracy</th>
<th>Reduce Costs</th>
<th>Streamline State and Federal Reporting</th>
<th>Increase Data Use</th>
</tr>
</thead>
</table>
| Rules           | Clear Business Rules | • eScholar provides Data Element Best Practices that describe how data in each field is used, and aligns them with NCES Data Handbook Definitions, SIF Objects and required Federal EDEN data elements.  
• By forming the data foundation for the entire scope of data, the eScholar Complete Data Warehouse truly provides the one, integrated single version of the truth. | X                      |              | X                                     |                   |
|                 | Single Version of Truth | • The eScholar Complete Data Warehouse provides the ability for local data standards that are historically familiar to be used while recoding these to global standards that will support global comparisons.  
• The eScholar Complete Data Warehouse can integrate and compare data from external sources, such as postal address changes, district boundaries, and congressional districts and geo code data to augment and improve the accuracy of data being collected from district sources. | X                      | X           | X                                     |                   |
<p>| Alignment       | Recoding     | • The eScholar Complete Data Warehouse collects the lowest level of individual student, assessment, attendance and other data so it can be analyzed and corrected. Using other systems, errors can be detected in aggregate numbers, but can not be corrected except at the lowest level of detail. The eScholar Complete Data Warehouse provides the full capability. | X                      |              |                                        |                   |
|                 | External Data Sources |                                                                                                                                         |                        |              |                                        |                   |
| Detail          | Lowest Level of Granularity |                                                                                                                                                                                                                                                                         |                        |              |                                        | X                 |</p>
<table>
<thead>
<tr>
<th>RADAAR Category</th>
<th>Sub-Category</th>
<th>Best Practice</th>
<th>Increase Data Accuracy</th>
<th>Reduce Costs</th>
<th>Streamline State and Federal Reporting</th>
<th>Increase Data Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accountability</td>
<td>Clear Auditability</td>
<td>• eScholar provides the capability to record, audit and analyze everything that happens to each record of data from the time it is collected until it appears in a report or data submission. This provides the ability to clearly track all changes and corrections in data.</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Defined Feedback Loop</td>
<td>• The eScholar Complete Data Warehouse provides the ability to collect and cleanse data from users that input the data (data stewards) and then return data to those end users in value-added reports. By providing this value-added feedback, eScholar dramatically increases the incentive for the data stewards to collect data correctly the first time.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Clear Sign-Offs</td>
<td>• The eScholar Complete Data Warehouse web interface provides the ability for the data stewards to evaluate the quality of the data and then accept the data as accurate, thus improving the credibility of the data.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Executive Leadership</td>
<td>• eScholar and its extensive user community have developed leadership best practices that help districts put in place the necessary responsibilities and leadership roles that will increase and continuously improve data quality.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Automation</td>
<td>Single Pass Error Catching</td>
<td>• The eScholar Complete Data Warehouse eTL is designed to catch as many errors in a data record as possible in the first check of that record. For example: If a record has an error in both a student’s birthday and enrolled school, the eTL will catch both, the first time the record is submitted. This provides the Data Steward the ability to correct all errors the first time and avoid the effort of having to go back and correct the record again later.</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RADAAR Category</td>
<td>Sub-Category</td>
<td>Best Practice</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>--------------</td>
<td>---------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Reporting      | Density Analysis | • The eScholar Complete Data Warehouse provides reports that make it clear how completely populated the data warehouse is, as well as, the frequency of values in each field. This powerful capability helps users to understand what data elements are available and how they can be used in analysis.  
• The eScholar Complete Data Warehouse web interface allows data stewards to understand the complete context of the data and consequently correct it with that data available. For example, data stewards can see the birth date and enrollment history data at the time they are correcting the child’s grade level. |
|                | Corrections in Context | X          | X | X |
Workplan

By using the RADAAR approach and following the six step workplan outlined below, districts will be able to successfully implement a data warehouse.

Data Warehouse Implementation

|---------|-----------------|--------------------------|---------|----------|-------------------|
| • Prioritize  
  – AYP  
  – Student Profile  
  – Assessments  
  – Research  
 • Develop Project Plan  
  – Data Team  
  – Policies  
  – Training  
  – Communication  
  – Incentives | • At Lowest Level  
  • Align with NCES Standards  
  • Use Open Standard Data Interfaces  
  • Look to SIF | • Use Artificial Intelligence  
  • Leverage Industry Standards  
  • Maximize Data Scope  
  • Buy an Integrated Data Model  
  • Ensure Referential Integrity  
  • Secure Data | • Benchmark Data Use and Cleanliness  
  • Use Artificial Intelligence | • Create Feedback Loops  
  • Continuously Improve |

1. Plan: Prioritize data needs and develop a project plan

Most districts are focused on complying with No Child Left Behind, including reducing achievement gaps among student groups and generating their Adequate Yearly Process (AYPs). Districts are also trying to give principals and teachers increased access to critical student demographic, assessment, disciplinary, and programmatic data to implement district strategies to improve student performance. Many districts are starting to use assessment trend analysis to improve instructional practice, such as the work being done with the Harvard Graduate School of Education and Boston Public Schools as defined in the recent book “Data Wise: A Step-by-Step Guide to Using Assessment Results to Improve Teaching and Learning.”

A useful approach to prioritizing desired data is to generate a Return on Use (ROU) analysis. Conducting an ROU evaluation requires the district to assess each data element by its usefulness and ease of collection relative to the other data elements. An illustrative result of this analysis is depicted in the table below.

---

Once the desired data are defined, create a project plan identifying the tasks, resources, deliverables, and timeframes required to implement the data warehouse. The project plan should clearly articulate the people and process for when and how data should be verified, collected, and loaded into the data warehouse.

Successful data warehousing districts develop a Data Team which is responsible for managing, verifying, correcting, and supporting district data. Critical roles include:

- Vocal, ongoing senior leadership support that clearly articulates the importance of the data warehouse
- A project manager who will oversee the initial implementation and provide a centralized point of accountability for ongoing collection of district data and adherence to data collection policies
- A data analyst who can translate educational policy data needs into identification of the individual data elements required to assess programmatic effectiveness
- A database administrator who will manage the consolidation and transmittal/loading of data files into the data warehouse
- Data source guardians who verify individual records by school

Another important component of the plan is training and communication. Districts that do not invest in training staff on data collection policies (or communicate the importance of data) have a harder time maintaining data integrity. Consider what human resource processes should change to support adoption of the data warehouse.
2. Capture Data

Your District should capture data at the lowest level of granularity (e.g., an assessment’s item level analysis by student), to maximize data usefulness and your ability to verify it. Districts should also choose a data warehouse that provides:

- Clear data definitions that continually evolve with NCES standards, enabling better alignment with state and federal reporting
- Maximum flexibility to capture all of your data needs, e.g., an Excel spreadsheet or handwritten list of students participating in choir
- Open standard data interfaces that provide domain specific templates defining what the data needs to look like to go into the data warehouse cleanly. You will need a solution to seamlessly deal with change, such as a new Student Information System (SIS).
- SIF compliance - SIF is a K-12 industry group that is working on technology (XML standards) to improve accuracy and reduce maintenance costs of data warehousing by focusing on the mechanics of getting data into the warehouse and keeping it clean. For example, when a student’s address is changed in the SIS, SIF agents can update data in the rest of the warehouse.

3. Extract, Transform, Load

Once your District has captured its data into a standardized format through the data interfaces, it is ready to be loaded into the data warehouse. Many Student Information System vendors and third-party intermediaries have built utilities to automatically bring your data into the warehouse. Data integration and cleansing (sometimes called ETL for extract, transform, and load) is the engine of data warehousing functionality. Districts need powerful and well-documented data transformation, integration and cleansing routines to reduce staff workload and improve data accuracy. These routines should introduce artificial intelligence to data cleansing, such as programs that help organize work, detect data outside of the norm or specific patterns, and flag these instances for human investigation. Routines should:

- Assist your District manage the complex workload of uploading multiple files throughout the year
- Verify data against district data standards and translate/recode data to national standards
- Catch most errors in a single pass and generate data error logs, such as duplicate records, fields with non-matching or “outlying/unique” values, that are too long, or with the wrong data type, and missing data (e.g., loading attendance data without a student or a student’s grade without a course)
- Produce status reports for district administrators including the percentage of desired fields populated
- Produce data investigation reports requiring further inquiry, such as birthdays, ethnicities and gender that change over time
- Synchronize data changes

4. Store
Your District needs a high performance data-model that is highly responsive to District queries and analyses and is designed to maximize retrieval speed and data accuracy. Data warehouse providers have spent years developing their data models which are intellectual property under copyright protection. Important components to assess when reviewing data models include:

- Scope of data. Data models should already include at least 80% of your data needs and be flexible and customizable to accommodate your remaining needs.
- Integrated data model. The data model should seamlessly integrate all your data domains, allowing analysis among multiple locations, assessments, teachers, students, programs, and costs.
- Referential integrity. Referential integrity ensures that once your data are integrated in the data warehouse, if one source changes, then this change can easily and currently be reflected throughout the database. Referential integrity also streamlines research requests. For example, some data warehouses are designed with each assessment in a separate area of the data warehouse. This is a design flaw that will limit a district’s ability to quickly and easily compare student achievement across assessments in a common skill set (i.e., fractions). Well designed data warehouses put all assessment data in a single area to increase data integration and comparability.

5. Report

Over time, districts find they spend less time cleaning their data. However, there is an initial upfront data cleansing process required that will pay dividends later on if done thoroughly. After the initial data load, districts should run baseline reports on the percentage of desired data elements collected and the quality of this data. Although every client has a unique scope of data needs, current e-Scholar state and district clients’ accuracy of data depends on data type. The chart below shows first load data accuracy rates by data type for a sample of eScholar clients.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Accuracy Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student ID</td>
<td>85%</td>
</tr>
<tr>
<td>Student Attendance</td>
<td>80%</td>
</tr>
<tr>
<td>Student Grade</td>
<td>80%</td>
</tr>
<tr>
<td>Student Course Attendance</td>
<td>50%</td>
</tr>
</tbody>
</table>

Districts use eScholar’s data density, quality and investigation reports to focus their data cleansing efforts. Districts should also maintain statistics on district use of data warehouse reports to measure how data are being used to drive decision-making.

Artificial intelligence is the next frontier in data warehouse reporting. Soon K-12 data warehouses should be able to proactively identify trends in student performance and flag struggling students for staff and parental intervention. Another example is an item level
assessment report that shows the teachers who are able to teach a specific concept particularly well. These teachers can then be tapped to share best practices to improve instruction.

6. **Improve and Maintain** to Meet Changing and Expanding Demand

As shown in the diagram below, the core process of data collection involves a continuous feedback loop between the schools and the district. The Data Team must ensure that data reports are sent to the proper data source guardians for verification and data correction in the source system. One of the central tenets of collecting accurate verifiable data is to collect data at the most granular level and have the person most knowledgeable about that data be responsible for verifying and updating it. The data warehouse should allow production of online teacher level reports to enable teachers to confirm the students currently in their class. School administrative assistants play a critical role in updating student demographic data. The district bilingual and special education coordinators must work closely with the schools to ensure proper identification of children with special needs.

![Continuous Feedback Loops](image_url)

Collecting, analyzing, and maintaining district data is an ongoing iterative process. The initial implementation of a data warehouse takes from three to nine months. However, the data warehouse is a living, constantly expanding vehicle that will continue to take the district to new places, constantly broadening its horizon. It provides the means to ask new questions and discover new frontiers. Ongoing maintenance of the data warehouse requires continuous attention to the details of collecting important granular data on all students. A strong project manager is required who will continuously monitor data quality and adherence to data collection policies.

**Moving Forward**
Although implementing a data warehouse can seem a daunting undertaking, your District will increase its ability to save staff time through streamlined reporting and improve performance through better understanding of what is working. Think of where our national public highway system was 50 years ago and where it can take us now. Similar benefits are waiting to be realized for the K-12 data infrastructure.