

The Higher Ed Data Cleanup Checklist

Data is valuable only when it can be trusted. Messy data, which can show up in errors such as missing fields or incomplete descriptions, can lead to reporting errors and reduced confidence in institutional insights. A data warehouse, such as Edify, helps institutions improve data quality by bringing information from multiple systems into a centralized environment, thus making it easier to identify gaps, standardize definitions, monitor data health, and resolve issues before they impact reporting, analysis, and decisions. By creating a single source of truth, a data warehouse helps ensure data remains accurate, reliable, and ready to support strategic decisions.

Share this infographic with the teams responsible for managing and reporting on institutional data to help build a common understanding of how messy data occurs and how a modern data warehouse can help create a trusted foundation for decision-making.



What is messy data?

Messy data indicates problems with the data itself, as well as with its documentation. This could look like:

- **Missing information**
- **Incomplete information**
- **Incorrect information**
- **Missing descriptions**
- **Misrepresented or incorrectly formatted** (e.g., a GPA field that says “four” instead of “4.0,” or percentages sometimes represented with the word “percent” and sometimes with the “%” symbol)

How does messy data come about?

HUMAN ERROR

- Required form fields are left blank
- Data is entered incorrectly
- Data entry practices are inconsistent
- Standard data entry guidance is lacking

DATA LOAD FAILS

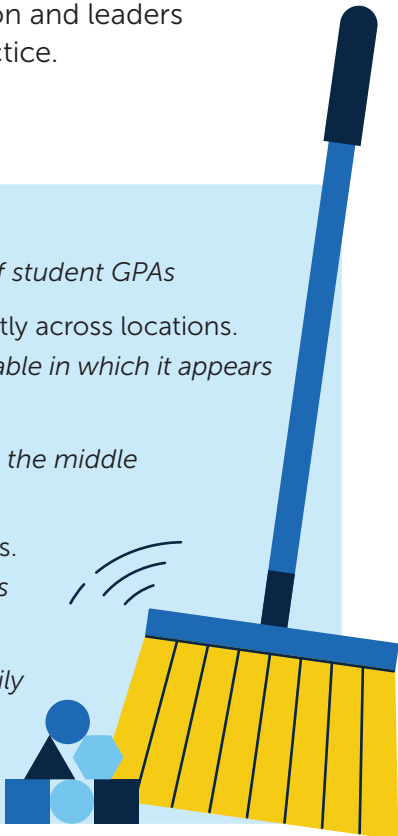
- Data fails to feed from one source to another

What is clean data?

Clean data is trusted, usable, and up to date. These data health principles help ensure that data is reliable, consistent, and ready to support confident decision-making. When institutional data is clean, staff spend less time gathering and validating information and leaders are equipped to put the most updated insights into practice.

DATA HEALTH PRINCIPLES

- ☐ **Completeness:** Data fields are fully populated
Example: There are no blank or null fields in a table of student GPAs
- ☐ **Consistency:** The same data is represented consistently across locations.
Example: A student’s ID number is the same in every table in which it appears
- ☐ **Accuracy:** The data is correct
Example: A student’s middle name in the SIS matches the middle name on their driver’s license
- ☐ **Format:** Data follows established formatting standards.
Example: All dates are represented as date/time values
- ☐ **Time Frame:** The data is fresh and up to date
Example: Class registration numbers are refreshed daily
- ☐ **Validity:** The dataset follows established rules
Example: All GPAs fall between 0 and 4.0



✓ **Together, these principles create a trusted foundation for reporting, analytics, compliance, and strategic decision-making across campus.**

Three ways a data warehouse facilitates data cleanup:

The process of bringing all your institution’s data together when implementing a data warehouse can surface hidden data problems or bring data cleanup needs to light. Once those problems have been addressed, your data warehouse can enable you to monitor data quality continuously and flag issues.

Here’s how a data warehouse can help improve data quality:



1 DATA DESCRIPTIONS

Maintain clear explanations that help both users and AI models understand what the data means, where it originated, how it was transformed, and how it should be used.



2 DATA QUALITY VALIDATION

Apply rules and automated checks to ensure that data meets set requirements before it is used.



3 DATA PROFILING

Ensure that data is populated, valid, and up to date.



How Edify helps you keep your data clean

In addition to these core capabilities, Edify, EAB’s higher education data management platform, includes tools that help institutions monitor and maintain data quality over time.



DATA FRESHNESS VALIDATIONS

Allow users who are delivering data to downstream systems such as Canvas to create validations on a delivery, ensuring that the intakes and builds associated with the delivery are refreshed within a specific time frame.



DATA PIPELINES

Allow users to link jobs together, define dependencies, and run multistep processes. Each run is tracked from start to finish, giving you clear visibility into progress and results.



JOB STATUS MONITORING

Allow users to view the details of each job—including start and end times, status, and messages—for processes such as extracting data from the LMS, loading it into Edify, and transforming and exporting it.