Quality Data — An Improbable Dream?

A process for reviewing and improving data quality makes for reliable — and usable — results

By **Elizabeth Vannan**

In partnership with British Columbia higher education institutions and the Centre for Education Information (CEISS), the BC Ministry of Advanced Education is implementing 22 data warehouses to collect standard data for reporting and analysis purposes. I manage the CEISS team responsible for implementation of those warehouses. A major challenge has been ensuring that the data submitted are of sufficient quality for accurate government reporting and to support decision-making. The institutions participating in the project are diverse, have varying business practices, and use more than five different administrative systems.

To address the data quality issues (inherent in any data warehouse implementation), the project team has developed an extensive process for reviewing and improving data quality. The methodology includes a thorough business practice review, an analysis of existing data quality, cleansing of existing data, and implementation of business practice changes aimed at improving the quality of data captured by each institution.

What Is Quality Data?

Quality data don't have to be perfect, just accurate, complete, consistent, timely, and flexible enough to meet business needs.

- Accurate — the data are free from errors.
- Complete — all values are present.
- Consistent — the data satisfy a set of constraints and are maintained in a consistent fashion.
- Timely — the data are available when required.
- Flexible — the data definitions are understood so that data can be analyzed in a number of ways.

Therefore, when establishing data quality standards, it's vital to understand the end use of the data.

The Cost of Poor-Quality Data

As more higher education institutions implement integrated administrative systems and data warehouses, they encounter the high costs imposed by poor-quality data.

Business Process Costs

When inaccurate information is entered into an administrative system, standard processes fail. At higher education institutions, process failures due to inaccurate information range from students not being registered for the correct courses, to inaccurate tuition billings, to payroll errors.

Cost to Rework Information

When data is entered incorrectly or is considered unreliable, data users must

- Collect the same information from other sources.
- Correct errors.
- Verify the data's accuracy.

All of these activities consume staff time and add cost.

Lost and Missed Opportunities

As the higher education sector becomes more competitive, institutions need quality data to support decision-making and assist in delivering services to students. Poor-quality data can result in substandard customer service (for example, potential students not receiving mailings due to inaccurate address information). Poor-quality data can also result in ineffective decision-making and potentially a loss of reputation.

Improving Data Quality

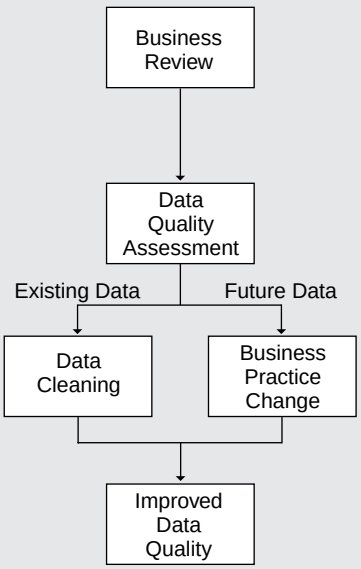
The first step in improving institutional data quality is realizing that problems exist. Data quality issues usually become evident when an institution starts planning a data warehouse project. For many institutions, such projects force them to start reviewing data and addressing quality problems.

To improve data quality, an institution must be prepared to

1. Take a hard look at their business practices related to the collection and recording of data.
2. Establish a data custodian (owner) for every piece of information they collect and make that custodian responsible for their data quality.
3. Perform a thorough and objective analysis of existing data quality.
4. Correct existing data quality problems.
5. Make permanent changes to business practices to improve future data.
6. Make the review and improvement of data quality an ongoing process.

Figure 1

Data Quality Improvement Process



As part of our own Data Definitions and Standards (DDEF) Project, the project team has developed a four-step process to help institutions identify data quality issues and establish an ongoing data quality improvement process. (See Figure 1.)

Business Review

The first stage in the project's process for improving data quality is to perform a thorough review of the institution's business practices related to the collection and recording of data. The review process asks

- How and when are data collected?
- Where and how in the institution's administrative system(s) are the data recorded?
- Are data recorded in more than one system? If so, how are they reconciled?
- What quality checks ensure accurate data capture?
- Who is the custodian of each data element?
- Who creates the data? (Who collects and enters data?)
- Who uses the data?
- What types of reporting are required?

The project involves as many stakeholders from the institution as possible. For the student subject area this typically includes the offices of the registrar, institutional research, IT department, and any other offices responsible for creating or using data recorded in the student registration system.

The results of the business review include

- A comprehensive understanding of the institution's business practices that impact data quality.
- Identification of individuals responsible for data collection and quality control.
- Identification of all data users and their requirements.
- Preliminary metrics to use in a quality assessment.
- Initial identification of problem business practices.

The business review can take anywhere from several days to several weeks, depending on the complexity of the data set being considered and the number of stakeholders involved.

Data Validation and Quality Assessment

Once the business review has ended, the next task is to assess existing data quality and completeness. A representative sample of data is extracted from the administrative system and placed in a staging area where it can be reviewed using a variety of software tools. The DDEF Project currently uses Oracle Discoverer for data validation and quality assessment. We build a series of queries to view every data element and compare the results to the established metrics. We found the software easy to learn and inexpensive, and it integrates well into our Oracle environment. A number of other software tools automatically perform data validation, but we found their cost prohibitive in a higher education environment.

The data validation and quality assessment process

- Establishes metrics to assess the validity of the data extracted from the administrative system. The metrics may include
  - Acceptable date ranges for elements such as birth date, registration date, and achievement date.
  - Acceptable syntax and values for elements such as course codes and program codes.
  - Estimates of counts of total student registrations and student records.
  - Business rules with which the data must comply. For example, a course section start date must come before the course section end date.
- Systematically reviews all the data elements, considering factors such as range of values, number of null records, duplicate records, compliance with business rules, and inaccurately recorded information.
- Provides a summary of data problems and a strategy for data cleansing.
- Tracks data problems to their source.
- Makes recommendations for business practice changes to improve data quality.

The data quality assessment can be a humbling and time-consuming experience for an institution, particularly if it's the first time they've looked critically at their own data. Typical problems found with student registration data include

- Students more than 2000 years old.
- Students not yet born.
- Course sections that took place before the college was established.
- Course sections that ended before they started.
- Registration clerks who "make up" program, course, and course section codes.
- Duplicate records, such as
  - Students with more than one ID number.
  - Students registered in the same course section multiple times.
- Inappropriate application of demographic flags
- Information provided by students but not recorded in the administrative system.

For every quality problem revealed, the institution identifies the source of the error, the magnitude of the problem (does it affect one record or thousands?), corrective action, and a strategy to prevent this error from occurring

again. Note that not all errors are preventable — simple keying errors are difficult to prevent, for example. Data will never be perfect.

Upon completion of the data validation and quality assessment process, the institution will have a complete picture of its current data quality and will have developed some initial strategies to resolve existing problems and prevent future ones.

Data Cleansing

Once data quality problems have been identified, the institution proceeds with data cleansing. Data cleansing goes to the source system and corrects errors and other problems identified during quality assessment.

In the DDEF Project, we prefer to do as much data cleansing as possible at the source — in the institution's administrative system. This ensures that the data is correct in the system of record and means that all future users of that data will have accurate information.

However, it's not always possible or practical to perform all cleansing in the administrative system. In these cases, the project also performs data cleansing in a staging area prior to its transfer into the data warehouse. Unfortunately, if data cleansing takes place after extraction of data from the source system, cleansing will have to be repeated every time the data is re-extracted.

Typical data cleansing includes

- Correcting data entry errors.
- Removing or correcting nonsensical dates.
- Deleting "garbage" records - records that don't contain valid data.
- Combining and/or deleting duplicate records.

Data cleansing is time consuming and therefore expensive. It's usually necessary to perform extensive data cleansing as part of implementing a data quality improvement process. Still, the ultimate goal is to reduce the amount of data cleansing required over time. Changing business practices, combined with regular quality reviews, should help limit the amount of routine cleansing required.

Changing Business Practices

During the preceding stages in the quality improvement process, an institution will have identified a number of business practices that negatively affected their data quality. The final stage in the project's quality improvement process involves

1. implementing changes to business practices that will improve data quality and
2. adopting an ongoing process to regularly review and improve data quality.

Implementing business practice changes within a higher education environment can be a challenge. It's important to involve the stakeholders who participated in the initial business review. It also helps to solicit executive support if the needed business practice changes are significant.

Business practice changes that improve data quality will range from simply reviewing data collection and entry practices with responsible staff to a complete reengineering of business processes. The types of business practice changes typically seen in our project include

- Educating data entry staff on the importance of accurate data entry.
- Updating code sets.
- Centralizing the creation of new codes.
- Ensuring that data is entered in the correct location in the administrative system.
- Consolidating data collection into one administrative system.
- Implementing validation routines within the administrative system.

For an institution to have continued quality data, it must also implement a regular process for reviewing and improving data quality. A continuing data quality process should

- Appoint someone within the institution responsible for monitoring overall data quality.
- Review the quality of data at regular intervals.
- Establish a regular updating cycle for all code sets.
- Provide a procedure for addressing data quality issues as they are identified.
- Establish data custodians who will be responsible for the quality of their data.
- Provide education on the importance of quality data to data custodians, data creators, and data users.
- Communicate data quality improvements to the users.

Quality Data Is Possible

The DDEF Project has had great success in assisting BC institutions to review and improve their data quality, largely from

- Educating institutions in the value of quality data and the high cost of inaccurate and incomplete data.
- Establishing a formal review and improvement process.
- Involving a broad group of stakeholders at each institution.
- Soliciting executive support to implement business practice changes.
- Making data quality improvement a regular process within each institution.

The project has collected data from higher education institutions for almost three years. In that time we've seen a vast improvement in data quality, thanks to establishing a consistent data quality improvement process. *e*

Elizabeth Vannan (evannan@ceiss.org) is Project Manager, Data Definitions and Standards, at the Centre for Education Information (CEISS) in Victoria, British Columbia, Canada.

Further Reading

Larry P. English, *Improving Data Warehouse and Business Information Quality* (New York: John Wiley & Sons, 1999).

Michael H. Brackett, *Data Resource Quality: Turning Bad Habits into Good Practices* (New York: Addison-Wesley, 2000).

Richard J. Orli, "Data Quality Methods," based on a public document prepared for the United States government, 1996 [http://www.kismeta.com/cleand1.html].