December 2003

Teaching GIGO: Data Quality in the Curriculum

Mary-Ann Robbert
Bentley College

Linda Senne
Bentley College

TEACHING GIGO:
DATA QUALITY IN THE CURRICULUM

Mary Ann Robbert
Bentley College
mrobbert@bentley.edu

Linda P. Senne
Bentley College
lsenne@bentley.edu

Abstract

High quality data and information are essential for decision-making; yet there is little focus on the inclusion of information quality in the current business curriculum. Quality issues are addressed in the introductory IS course but may not be seen again. This paper reviews current texts for inclusion of quality and notes little change in the model curriculums inclusion of quality. Suggestions for database assignments addressing data quality issues and an information quality component for a data warehousing course are included.

Keywords: Data quality, information quality, IS'97 model curriculum, IS 2002 model curriculum, database, data warehousing

Introduction

Data are commonly defined as distinct facts about a person, place or object that are in a format for manipulation by a computer while information is characterized as data put into human context. Courses on database management systems frequently stress topics on data quality from the perspective of data processing. The subject of information quality—the value derived from information products that provide managers with data for making better decision - is generally omitted from the material in DBMS courses.

Khalil *et al.* (1999) note that at most IS students are exposed to topics that impact information quality but are not provided with an understanding of the principles of data quality that are prerequisite to quality information. The factors that they consider to have the most influence on what students learn include

- college curriculum,
- the instructor and
- textbooks.

There has not been a significant change with respect to "quality," either information or data, from the IS'97 Model Curriculum (Davis *et al.* 1997) to the IS2002 Model (Davis *et al.* 2001). In the IS2002 Model Curriculum, there are two main places where the Model Curriculum addresses the topic of data quality, a term that is not specifically defined in the document. This paper examines possibilities for including information and data quality in these courses and elsewhere in the business school curriculum.

Based on our review of recent editions of typical textbooks, little has changed in the treatment of data quality over the past few years. Authors appear to cover data quality when they look at data from the user's perspective in order to discuss information in business systems. In textbooks that focus on developing and managing a DBMS, the authors write from an IS perspective. The word "integrity" appears more often than "quality." Thus students are frequently introduced to the topic of data quality in introductory MIS courses, but this focus on data quality is not reinforced in advanced courses. This paper provides examples from textbooks used at many business schools.

Given the current curriculum and typical textbooks, instructor may have to develop their own modules for examining the subject. The paper suggests information quality modules and assignments that can be included in database management and data warehousing courses.

Study of Information Quality

Information quality is more than just data integrity and accuracy. Wang and Strong (1996) originally identified four dimensions of data quality and fifteen measurable attributes

- Intrinsic data quality: accuracy, believability, objectivity, reputation
- Contextual data quality: value-added, relevancy, timeliness, completeness, appropriate amount of data
- Representational data quality: interpretability, ease of understanding, representational consistency, concise representation
- Accessibility data quality: accessibility, accessibility security

A subsequent paper further develops the attributes with a focus on information quality and derives a methodology for information quality assessment (AIMQ) on these attributes (Lee *et al.* 2001). The fifteen attributes are used in this paper as the standard for complete inclusion of information quality attributes in CIS courses.

In addition to the characteristics that have been defined to determine information quality, a number of authors recommend quality models. Dvir and Evans (1996) and Becker (1998) propose a TQM framework that allows classical statistical quality control techniques like Pareto and control chart to be used. These techniques can be applied to measuring, tracking and improving data quality. Khalil *et al.* (1999) have proposed an information quality model based on total quality management and marketing. Their matrix compares product and service quality with conformance to specification and meeting customer, i.e. user, expectations.

Although attributes like those defined by Wang and Strong appear to define information quality clearly, users and IS professionals interpret them differently because the two groups relate to data differently. For the former data are facts that correspond to objects and events in the real world; for them data is concrete. For IS professionals, however, data is more abstract; it tends to be the bits and bytes representing what is defined in the data dictionary and processed by the application. Since most of the attributes defined by Wang and Strong (1996) are obvious to business professionals who regularly deal with data; little training is required to convince them they need quality data, and they view the IT staff as the group responsible for delivering quality data to them. Because the IT staff's perspective is different, however, they tend to overlook many aspects of data quality that users require for information quality.

According to Becker (1998) the research strongly suggests that information system professionals should deliver more than accurate and objective data to the users. Seven common data quality problems seen by end-users include: data corruption due to incorrect conversion, historical and current data having different meanings, the same data having more than one data definition, missing data having different meanings, the same data having more than one data definition, missing data, hidden data, missing granularity, and violation of integrity rules (Mathieu, Khalil 1998).

By introducing appropriate modules in database management courses, instructors can give future IT professionals greater understanding of the user perspective on data so that they, the professionals, understand the expectations of their customers, the users of the data. Such modules can help students prepare to deal with total quality management issues after they have left school.

Strong suggests focusing on three areas to increase student's knowledge of data related issues – the role of information in organizational decision making, how information flows in an organization, and the dynamics of data (Strong 2002). Redman (1996) examines quality issues and suggests a data policy that covers security, privacy and rules of use; inventory of data assets; data sharing and availability; data architecture; planning; and the role of quality.

Information Quality Education

Students can learn about IQ as part of an undergraduate degree program, a graduate degree program, a certificate program or vendor provided seminars. This paper focuses on learning information quality concepts within a college degree program.

A separate course on information quality can be offered in the general business curriculum or specifically in the information systems department. Craig Fisher (2002) offers a course, Data Quality in Information System, at Marist College. This course is an IS elective that was introduced in 2002 after being offered as an experimental course. Chippewa Valley Technical College offers two management courses, Management: Control and Quality Management, which has a TQM focus, and Project Management, which deals with quality processes of project quality planning, quality control, and quality assurance. The

Continuing Education Program at University of Alabama, Huntsville, includes a course, Quality Management and Performance Measurement that offers a foundation in the PMI® Knowledge Area of Project Quality Management. The University of Missouri, St. Louis, offers Quality Management 430, an applied course on total quality management. The University of Melbourne undergraduate management degree with a specialization in Operations and Strategic Management, and their Graduate Diploma in Management Studies require the Quality Management course. This course provides an introduction to the major theories and models in quality management.

In lieu of or in addition to a separate course, data quality and information quality concepts can be included within current courses, not as an aside but as a fundamental requirement for valid information. Marketing courses focusing on customer satisfaction or CRM should include quality as should management courses dealing with decision- making. Data quality should also be an integral part of the database course. Mathieu and Khalil (1998) noted that students do not universally receive instruction on the overall importance of data quality in the design and implementation of databases. Though topics such as data integrity, security and concurrency control are usually covered they are not tied into general data quality.

We examined the model curricula to determine how much focus is placed on information quality. In their 1998 paper, Mathieu and Khalil (1998) reviewed the IS'97 Model Curriculum and the IRMA/DAMA Curriculum Model. They noted that the IS'97 Model focuses primarily on "principles of quality improvement" and on "software quality." There is only one mention of information quality in IS'97, as a topical area in IS'97.8. The IRMA/DAMA Model lists quality control and information quality as an organizational issue in an Information Technology course and as a topic in Information Resource Management Principles.

The IS2002 Model Curriculum (Davis 2001), under Information Systems Theory and Practice, states that students will be introduced to concepts and theories that explain or motivate methods and practices in the development and use of information systems in organizations. The concepts and theories will include systems, management, and organization, information, **quality**, and decision making. The Information Systems Deployment and Management section maintains that management of the information systems function, systems integration, and project management to ensure project **quality** are integral components of this curriculum area. The model suggests that quality should be introduced in IS 2002.1 – Fundamentals of Information Systems with quality of information listed as a topic. The discussion section of IS 2002.8 – Physical Design and Implementation with DBMS mentions that **quality** assurance measures implemented as project standards will be used to control project **quality** and risk. Quality assurance is also mentioned under Learning Unit Number 127 for this course and in IS 2002.9 – Physical Design and Implementation in Emerging Environment. IS2002.3 Information Systems Theory and Practice mentions quality.

In a basic MIS text, **Introduction to Information Systems**, 9th edition, James O'Brien (2000, p 29) defines the "attributes of information quality" in three-dimensions – time, content, and form. The attributes of time are timeliness, currency, frequency, and period. The dimension "content" has attributes accuracy, relevance, completeness, conciseness, scope, and performance. The "form" dimension includes clarity, detail, order, and media. These attributes align themselves well with those proposed by Wang and Strong (1996). Ronald Thompson and William Cats-Baril's **Information Technology and Management**, 2nd edition (2003) contains material similar to that in O'Brien. The authors discuss relevance, timeliness, accuracy, reliability, completeness and granularity in *Chapter 6 Information Systems and Organizational Competition* as desirable characteristics of information but do not mention data here (p.206). They also allude to accuracy and completeness under the topic of maintenance in *Chapter 11 Information Systems Development* (p. 388). A similar book**, Management Information Systems for the Information Age 3rd** edition by Stephan Haag, Maeve Cummings, and Donald J. McCubbrey (2002), talks about information flow but neither "quality" nor "accuracy" is considered.

We can begin the examination of database management texts with C. J. Date's **An Introduction to Database Systems**, 7th ed, (2000). It covers integrity, concurrency and security but does not talk directly to the issues of data quality except under data cleansing. Here he notes that few data sources control data quality adequately and thus the data must be cleansed. The authors also looked at **Database Processing** 7th edition (2000), by David M. Kroenke, a text appropriate for MIS majors. Kroenke discusses "data integrity," which he defines as logical consistency in the introduction and under Sharing Enterprise Data. Data accuracy is discussed as a need for data warehouses. Peter Rob and Carlos Coronel in **Database Systems: Design, Implementation, & Management**, 4th edition (Course Technology, 2000) introduce data redundancy in the first chapter where they discuss data inconsistencies and accuracy. Data integrity issues are examined under database integrity management and transaction management and concurrency control. Under database administration they note that, if the information is accurate and timely, its use is likely to trigger actions that enhance the company's competitive position and generate wealth. Silberschatz *et al.* (2002) in **Database System Concepts,** 4th Edition, handle quality issues indirectly in a chapter, Integrity and Security. Though the above authors deal with some attributes of information quality, they do not mention information quality by name nor discuss it as an issue - at least not in the index.

Integration in Current Courses

Before curriculum changes have been introduced, information quality can be contained within current courses. Finding students lacking in knowledge of the overall role of information quality in the design and implementation of databases, data warehouses and information systems in general, I teach the basic concepts in my courses. I stress information quality in projects, homework, and discussions in my database and data warehousing courses.

Frost argues that students graduate without being able to solve business problems using database technology and recommends placing students in teams and have them design solutions to real business problems (Frost 1997). I agree and contend that information quality must be a fundamental part of the solution.

Project for a Database Course

I have experimented with team projects in the undergraduate and graduate database management courses. I was most successful when the project was broken into modules that were collected and reviewed throughout the semester. I then developed three assignments where the essence of information quality appears to the students through the data itself.

Students individually create a simple database (we use Oracle). I have been most successful with a five table orders database containing customers, orders, order lines, sales rep and products. I post the data for the tables in a text file on the web. Students must use the data given, but I do not describe or name the fields. Students must add five new customers, two new products and five new orders. This assignment is just graded on completeness and is worth a minimal percentage of their grade.

The next week the students are informed that their individual company has merged with three other companies and reports must be generated from the aggregate data. I have not found any differences in assigning groups, randomly selecting groups or having students select their own group. The groups get together in class and exchange user names to grant read privileges to teammates. The assignment is to individually generate reports (list of all customers, list of products sold in 2001, itemized orders for past 5 years with totals per order and per customer, etc.) for the corporate office containing all the information but without duplicates. Since this is an individual assignment, only the data is shared. Because the companies are newly merged, data must remain as is until a decision is made on if and how the data should be integrated. No data may be changed by anyone and no collaboration between team members is permitted. Students are given two to three weeks to complete the assignment.

The class after the assignment is given is dynamic with students discovering for themselves data quality issues. Representational data quality issues dominate. Students are amazed at the total lack of representational consistency. Some have first and last name as one field, others two; some have address as one field, others many; and some have made up extra attributes. Completeness is exemplified through the orders lines. Many students added orders but never purchased anything, i.e., there were no corresponding order lines. Next, the intrinsic data quality issues, especially accuracy, are discussed. There are any number of typos, misspellings, extra spaces and completely unbelievable data. Students are now ready for a discussion of data quality and information quality. The information in their reports must be accurate yet they only have the data provided. At this point we discuss possible ways of dealing with dirty data in historical databases and in data that needs to be integrated.

When the assignments are submitted, I grade them individually, but I group the papers by team. If the answers for all four member of the group do not agree, I subtract 10 points from each report with differing totals and require the group to determine who has the correct answer. Students with correct answers get the points returned. This exercise demands that the students reexamine the data quality issues. I have never found the error caused by an SQL coding error. Discussion now focuses on validity of solutions used to integrate the data such as what is a duplicate and what is a reasonable way to handle each record. For example, duplicate customers can be defined differently from duplicate orders. At the end of this session students usually determine that, if they had merged all the data, it would have been much easier.

In assignment three, the company has decided to form one centralized database that will contain all the data currently available plus new entries. Students work as a team to create a new database that addresses all the data quality issues. At this point all the dimensions and attributes of data quality as discussed above are covered. Although there are time constraints, many standard database topics such as integrity and security can be woven in. Rather than just identify quality problems at this point, students must define rules to prevent the problems from occurring when migrating the data and when adding new records. They need to handle the original orders that appeared identical but were placed at different stores as well as, different products with the same

names and same ids, and products with the same names but different ids, products with different names but the same ids. The warehouses for each company have the same id and the sales reps have the same id plus commissions.

A final reflection requires students to determine the advantages and disadvantages of storing the data separately or together. This can be done as a class discussion, written paper or exam question.

Quality Component for a Data Warehouse/Data Mining Course

We offer a graduate level course in Data Warehousing/Data Mining. This course contains a module on information quality in addition to coverage of data quality throughout other topics. The course description and syllabus are available at **http://blackboard.bentley.edu/bin/common/course** (enter the site for CS753 as a guest). Students in the course complete a Perseus Survey on-line. This survey contains many of the topics used by Wang and Strong to determine information quality dimensions and attributes. Though the sample size, 35, is not really significant, the results summarize class views of quality prior to discussion. The initial lecture displays class results and compares these to those delineated by Wang and Strong. The class identifies and ranks the attributes.

We offer a graduate level course in Data Warehousing/Data Mining. This course contains a module on information quality in addition to coverage of data quality throughout other topics.

Students in the course complete a Perseus Survey on-line. This survey contains many of the topics used by Wang and Strong to determine information quality dimensions and attributes. Though the sample size, 35, is not really significant, the results summarize class views of quality prior to discussion. The initial lecture displays class results and compares these to those delineated by Wang and Strong. The class identifies and ranks the attributes.

I then present a lecture on quality that contains information about data quality, repairing dirty data and changing the process to prevent the entry of bad data. Reliability of the results obtained from the data warehouse is stressed. I assign readings on a selection of four to six current papers in the literature, distributed randomly; each student has two papers to read. At the next class we discuss the different perspectives followed. When a suitable case is available and time permits it is included.

I have given large data sets and asked the students to clean the data using an Extract, Transform and Load (ETL) tool. We have one copy of Cognos in the lab but students were encouraged to test tools available for download on the web. Students were shocked that they had to write the rules. They thought you pushed the button and out came clean data. This assignment was interesting but lengthy.

Students are required to design and implement a data mart as part of the course requirements. In their documentation they must describe how they ensured data quality at each step of the process. Students must estimate the information quality of results obtained from their data mart. The goal of meeting the users' needs with high quality information is stressed. Projects are reviewed across the information quality attributes.

Conclusion

The explicit study of information quality is essential for those managing data and those relying on data for decisions. In today's business school, it is not sufficient to hand wave and say the material is covered in all courses. The curriculum needs to be examined and the flow of quality concepts learned by all students needs to be examined. An elective course that examines the quality issue in depth is excellent but, if it is an elective, it is insufficient. Knowledge of information quality must be a goal of the curriculum not only for Information Systems majors at a minimum but also for all business students as well.

References

Becker, S. "A Practical Perspective on Data Quality Issues," *Journal of Database Management*, (9), 1998, pp. 35 – 37.

Curriculum Model 2000 of the Information Resource Management Association and the Data Administration Managers Association, **http://www.irma-international.org/downloads/pdf/irma_dama.pdf**.

Date, C. *An Introduction to Database Systems*, 7th ed,, Boston, MA: Addison-Wesley, 2000.

Davis, G., Gorgone, J., Couger, J., Feinstein, D., Longenecker, Jr., H.(1997). *IS'97 Model Curriculum and Guidelines for Undergraduate Degree Programs in Information Systems,* **http://www.cis.usouthal.edu/faculty/feinstein/IS97/document/is97_title.htm**.

Davis, G., Gorgone, J., Couger, J., Feinstein, D., Longenecker, Jr., H.(2001), *Is 2002: An Update of the Information Systems Model Curriculum*, **http://www.spatial.maine.edu/SIEWEB/IS_ModelCurriculum.pdf**.

Dvir, R.,and Evans, S. "A TQM Approach to the Improvement of Information Quality," *Proceedings of the 1996 Conference on Information Quality*, Cambridge, MA, 1996, pp. 207 – 220.

Fisher,C. "A New Course: Data Quality in Information Systems," *Proceedings of the Seventh International Conference on Information Quality,* p. 203.

Frost, R., "Teaching Design to Solve Business Problems,"*Journal of Database Management*, (8) 1997, pp. 37 – 38.

Haag, S., Cummings, M., McCubbrey, D. *Management Information Systems for the Information Age* 3rd edition, New York: Irwin/McGraw-Hill, 2002.

IRMA/DAMA Curriculum Model (1996), **http://www.irma-international.org/crclm919.html**.

Khalil, O., Strong, D., Kahn, B., Pipino, L. "Teaching Information Quality in Information Systems Undergraduate Education*," Informing Science*(2:3), 1999, pp. 53-59.

Kroenke, D., *Database Processing* 7th edition, Upper Saddle River, NJ: Prentice Hall, 2000.

Lee, W., Strong, D., Kahn, B., Wang, R. "AIMQ: A Methodology for Information Quality Assessment," accepted for publication *Information and Management*, 2001.

Mathieu, R. and Khalil, O. "Data Quality in the Database Systems Course," *Data Quality Journal* (4:1), Sept. 1998.

O'Brien, J., *Introduction to Information Systems*, 9th edition, NY: Irwin/McGraw-Hill, 2000.

Redman, T., *Data Quality for the Information the Information Age*, Boston, MA,: Artech House, 1996.

Rob, P. Coronel, C., *Database Systems: Design, Implementation, & Management*, 4th edition, Boston, MA: Course Technology, 2000.

Silberschatz,A. Korth, H. and Sudarshan, S. *Database Concepts*, 4th Ed., Boston, MA: McGraw Hill, 2002.

Strong, D.(2002), "Recommendations for Information Quality Education," *Proceedings of the Seventh International Conference on Information Quality,* p. 202.

Thompson, R. and Cats-Baril, W., *Information Technology and Management*, 2nd edition, Boston, MA: McGraw-Hill, 2003.

Wang, R., and Strong, D., "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems,* (12:4), Spring 1996, pp. 5- 34.