Developing a Graduate-Level Course on Information Quality Tools: Closing the Gap between Theory and Practice

Teaching Case
Yinle Zhou
Information Science Department
University of Arkansas at Little Rock, USA
yxzhou@ualr.edu

John R. Talburt
Information Science Department
University of Arkansas at Little Rock, USA
jrtalburt@ualr.edu

Abstract
A discussion of the objectives and content of a graduate-level course on information quality tools that was designed to give students a practical, hands-on experience with the many of the tools and techniques currently used by information quality professionals. Because the course is embedded as a core requirement in two information quality degree programs, it has been offered on a regular basis and has evolved into its present form since its initial offering in 2006.

Keywords: information quality, data quality, information and data quality tools

Background
Information and data quality have their roots deeply embedded in the practical application of processes and policies that were developed to address the problems uncovered in enterprise data systems during the data warehousing movement of the early 1990s (Inmon 1992). It was then that most companies came face-to-face with the poor condition of their operational data systems that had traditionally been managed in separate silos across the enterprise. Data quality now extends beyond its original tactical focus on data assessment, cleansing, and monitoring, to the much broader view of information quality that encompasses strategy, policy, governance, culture, and business impact (Talburt 2011). Information quality (IQ) is now an emerging discipline concerned with maximizing the value of an organization’s information assets and assuring that the information products it produces meet the expectations of the customers who use them (Talburt et al. 2009). The evolution of IQ into a more theory-based, academic discipline can be seen in the introduction of information quality courses into existing curricula (Strong et al. 2005), the establishment of information quality degree programs (Informaiton Quality 2011), the recent launch of the Association for Computing Machinery (ACM) Journal of Data and Information Quality (JDIQ 2011), and the International Organization for Standardization (ISO) master data quality standard 8000-110:2009 (ISO 8000 2011). Furthermore, agreement on what comprises the body of IQ knowledge is beginning to stabilize around the research on IQ tasks, knowledge, and skills (Yonke et al. 2011) conducted by the International Association for Information and Data Quality in it development the Information Quality Certified Professional (IQCP) credential (IQCP 2011).
To address the growing demands by government and industry for trained professionals who understand the concepts, principles, tools, models, and techniques for information quality and who can solve current and future IQ problems, many colleges and universities are beginning to incorporate information quality courses and seminars into their curricula. At the Massachusetts Institute of Technology (MIT), the Information Quality Program (MITIQ) has been the leader in promoting and disseminating research in IQ through its sponsorship of the annual International Conference on Information Quality (ICIQ). Another milestone has been the creation of the first graduate degree program in information quality at the University of Arkansas at Little Rock (UALR) in 2006. Created through a joint effort of industry and academia, the curriculum strives to balance IQ theory with IQ practice. Beginning with a MS degree in IQ (MSIQ) in 2006, the program now includes a Graduate Certificate in IQ and an IQ track in the Integrated Computing Ph.D. program.

**MSIQ Curriculum**

The MSIQ program is based on two complementary philosophical approaches to learning: problem-based learning and foundation-building learning (Lee et al. 2007). The curriculum has been developed so that faculty can deliver a course in a way that maximally benefit from both approaches, with the understanding that some courses will focus more on one approach than the other based on the nature of the materials covered in the course.

The MSIQ program requires 33 semester credit hours comprising 9 graduate courses. Students must also complete a written thesis or an externally-sponsored IQ improvement project. The MSIQ curriculum is consistent with relevant guidelines provided by the Association for Computing Machinery (ACM) and the Association for Information Systems (AIS). Table 1 shows the curriculum requirements currently in effect (MSIQ 2011).

To increase access to the program, the UALR classes required for the MSIQ Program are webcast live through the Internet using Wimba Live Classroom available through the Blackboard Course Portal. This allows students in the program to participate in the classroom on-campus or online. In the case of major examinations, students whose place of residence is within 25 miles of the UALR campus (the definition of a local student) must take their major examinations in-person in the assigned classroom on-campus. Students whose place of residence is more than 25 miles from the UALR campus (the definition of a remote student) must arrange to take their major examinations under the supervision of an exam proctoring service approved by the program coordinator.

<table>
<thead>
<tr>
<th>Table 1. MSIQ Curriculum</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MSIQ Required Courses</strong></td>
</tr>
<tr>
<td>Information Science 5345 Information Visualization</td>
</tr>
<tr>
<td>Information Quality 7303 Principles of Information Quality</td>
</tr>
<tr>
<td>Information Science 7310 Information Systems Analysis</td>
</tr>
<tr>
<td>Information Quality 7367 Information Quality Policy and Strategy</td>
</tr>
<tr>
<td>Information Science 7320 Database Systems</td>
</tr>
<tr>
<td>Information Quality 7322 Information Quality Theory</td>
</tr>
<tr>
<td>Information Quality 7342 Information Quality Tools and Industry Landscape</td>
</tr>
<tr>
<td>One non-IQ specific elective such as</td>
</tr>
<tr>
<td>Information Science 5325 Data Mining Concepts and Techniques</td>
</tr>
<tr>
<td>Information Science 5330 Database Security</td>
</tr>
<tr>
<td>Information Science 5339 Network Security</td>
</tr>
<tr>
<td>Information Science 7325 Advanced Data Mining</td>
</tr>
<tr>
<td>Information Science 7330 Information Systems Security</td>
</tr>
</tbody>
</table>
IQ Tools and Industry Landscape Course Development

Course Objective

Because the students in the program come from a wide variety of backgrounds, the IQ tools course was put into place to assure that all students completing the program are exposed to some of the technical issues in data quality such as data assessment, analysis, standardization, cleaning, and enhancement. And even though many of the graduates from the program go into one of the many non-technical IQ roles such as IQ manager, data steward, or team leader, it is still important that every student has some understanding and appreciation of these fundamental data quality activities. This experience and understanding is especially important for those students going into roles where they will be involved in creating requirements for developing or acquiring IQ applications.

Course Design Development

The IQ Tools course has been offered in every Spring semester since 2007. The first IQ tools class was given just locally at UALR campus without Wimba Live Classroom® equipped. From the Fall semester of 2007, IQ program started to use Wimba for distance education.

IQ Tools lectures

To fulfill the objective of this course, the IQ Tools and landscape course has been designed to contain lectures and labs. One issue for the MSIQ program is the lack of textbooks specific to this area of study. In many cases the courses must rely on the many books produced by the IQ practitioners and consultants. Because the IQ tools course is practice-oriented, this does not present a problem. At first several offerings, the book Data Quality Assessment (Maydanchik 2007) has been used as textbook. This book focuses on how to assess data quality problems from introducing the general causes of poor data quality to designing and implementing data quality rules. Students gave feedbacks that this book is very practical and useful. Some of them introduced this book to their data quality team as a guide reference.

In recent offerings of the course, another book Three-Dimensional Analysis: Data Profiling Techniques (Lindsey 2008) has been added to the syllabus as another textbook. Data profiling is considered as the first and very important step for any data quality project. This book focuses on profiling techniques specifically and gives the idea of three-dimensional profiling data: column analysis, table analysis, and cross-table analysis.

In addition to the books used as IQ lecture materials, another lecture part is about introducing the landscape of IQ industry and this part is always up to date. Several guest speakers have been invited to
give talks about IQ topics in this class, like Dr. Richard Wang from MIT IQ program, Pam Needly from Rochester Institute of Technology, and so on. Students are always encouraged and enriched by their talks and opinions.

**IQ Tools Labs**

The books and other lecture materials supplement the laboratory portion of the course in which the students complete assignment using IQ related software. At beginning, only DataFlux dfPower Studio® was offered as IQ software.

DataFlux dfPowerStudio® is a commercial software system provided to the university through an academic license by SAS DataFlux®. SAS DataFlux is a recognized leader in IQ software and is consistently rated in the “Magic Quadrant” of the annual Gartner Research ratings of Data Quality Software companies. With this software, students can practice data profiling, standardization, integration, enhancement and many other IQ related functions. DataFlux dfPowerStudio® has been used as the main tool introduced in the lab part.

In the recent offerings, other software have been gradually added to the lab tools list, such as MySQL, R Statistical Package, Google Refine, OYSTER, and so forth. The following is what have been offered in 2011 Spring semester:

- **Database Tool:** MySQL (2 sessions)
- **Data Quality Tools:**
  - DataFlux dfPower Studio 8.2® (4 sessions),
  - Google Refine (1 session)
- **Entity Resolution Tool:** OYSTER (2 sessions)
- **Database Tool:** MySQL (2 sessions)
- **Statistical Tool:** R Statistical Package Version 2.8.0 (2 sessions)
- **Data Mining Tool:** WEKA data Mining GUI Version 3.6.0 (1 session)

In this semester, in addition to DataFlux dfPower Studio 8.2® continues to be introduced, there is another data quality software added: Google Refine. Google Refine is an open source system recently introduced by Google Lab.

DataFlux dfPower Studio® and Google Refine, both are data quality tools, but each has different features. As a commercial tool, DataFlux focuses on enterprise data management issues from managing data quality, data integration, and data enhancement to master data management (MDM). DataFlux gives students experience with a mature software package that provides a relatively complete solution. Students complete exercises related to each part of the DataFlux five-step solution that includes profiling, quality, integration, enrichment and monitoring (Data Flux 2011). Google Refine was only recently introduced is still under development as an open-source project (Google-refine 2011). Google Refine can perform data profiling, standardization, extraction, and other functions with Google Refine Expression Language and JSON.

Students are given assignments on different topics, such as data profiling, improving quality, data integration, data enrichment, etc. For each topic, a lecture and software demo are first shown in class. For example in data profiling, students are taught different techniques in class such as structure discovery, data discovery, and relationship discovery, and how to create and analyze different profiling reports. The exercises for each topic vary from simple, such as following the in-class demo template, to more challenging questions that require deeper analysis.

In recent offering of the class, the database tool MySQL was introduced for two sessions. Although all students in the program are expected to have had some database experience, MySQL is introduced as a common data source for the assignments. In addition to being given a refresher on basic database manipulations, students are also taught how to connect to MySQL from DataFlux as a starting point for several of the class assignments.

Data integration, especially entity resolution (ER), are critical data processes that can have a profound effect on IQ. ER is the process of determining whether two records in an information system referencing
real-world objects are referring to the same object or to different objects (Talburt 2011). ER can enhance data quality by providing the single view of an object in the entire information system. On the other hand, data quality can help with ER result by offering more accurate, complete data. OYSTER is an Open Source ER system developed by the ERIQ Research Center at the University of Arkansas at Little Rock (ualr.edu/eriq). OYSTER provides access to a variety of entity resolution algorithms that enables users to uncover duplicate and redundant entity references. In addition, OYSTER also provides support for entity identity information management (EIIM) and persistent entity identifiers (Zhou and Talburt 2011). Demos of how to configure OYSTER and run it in different modes such as identity capture, identity resolution, and identity update are given in class. Further explorations of ER and OYSTER are in course INFQ 7348 Entity Resolution and Information Quality offered in the Fall semester.

The R-Package is a powerful, open-source statistical language that is being widely adopted as an enterprise solution. It is used in the course to introduce students to statistical process control which supports process data quality management. Corresponding R-Package assignments are designed and given to students.

WEKA is a well-known, open source data mining tool developed by University of Waikato in New Zealand. Similar as ER, there is a close relationship between data quality and data mining. In this lab, students not only learn how to use WEKA to perform data mining tasks such as creating decision trees to classify and predict data, but also experiments that help them understand how data quality issues in source data can affect data mining results. Further discussions about data mining and WEKA are offered in data mining courses.

All of the tools are installed at a server which is located at UALR campus. Students can access the tools by Remote Desktop Connection which is a Windows-based product coming with every Windows Operating System. The Remote Desktop Connection is platform independent. Apple users can use it by installing Remote Desktop Connection Client. There are three reasons for choosing Remote Desktop as a remote access client for students. The first is the need to control the access and use of commercial software packages like DataFlux dfPower Studio® even when used under an academic license. The second is that it offers students a “one-stop” location where they can access all the tools they need for the class, even open source packages that students could install locally, to insure that everyone is using the same version. The third is that the server offers a larger working space and more processing power than most students would have on a local machine. This is especially important when working on challenge projects in a team environment.

**Team Project**

In addition to individual assignments, in the recent offerings, the course also includes a major team project called the “Data Integration Challenge”. In the challenge project, student teams are given several files containing customer information that has been extracted from a large, synthetic occupancy history file [3]. The extracted records are disrupted by creating duplicate records and also introducing data quality problems into individual records such as typographical errors, missing values, and common abbreviations or nicknames.

The student teams are formed and presented with the challenge early in the course and are asked to re-integrate the information and submit their results at the end of the course. Each team must develop its own strategy for solving the challenge and is free to use any of the tools from the course in the solution design. Because the customer information given to the students was extracted from a single, clean source, it is possible to measure the degree to which each team is successful in re-integrating the information (Zhou and Talburt 2010).

There are a number of ways to compare the similarity of grouping (partitions). One measure that was specifically developed to compare ER results and relatively easy to compute is the Talburt-Wang Index (TWI). (Talburt et al. 2007)

If A and B are two partitions (groupings) of a set S, and V is the set of overlaps between A and B (i.e. the set of non-empty intersections between groupings in A and B), then the TWI measure of similarity between A and B is defined as
The value of TWI will always be a value between 0 and 1, and its value will only be 1 when the two groupings A and B are identical. A smaller value of the TWI indicates less similarity between the groupings.

In the case, where A is the correct grouping and B is a team result, then the TWI value can be interpreted as entity resolution accuracy. The optimum result from ER will be when the groupings are identical giving a TWI value of 1.

In the latest offering, 2011 Spring semester, there were 28 students in class and they were grouped into 7 teams with 4 students in each. The Data Challenge project was assigned after mid-term exam. Three ER results were required to submit, two intermediate and one final. Table 2 shows the three results from the seven teams.

<table>
<thead>
<tr>
<th></th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>T6</th>
<th>T7</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>.3881</td>
<td>.5426</td>
<td>.2721</td>
<td>null</td>
<td>.3961</td>
<td>null</td>
<td>.4415</td>
</tr>
<tr>
<td>II</td>
<td>.4888</td>
<td>.5555</td>
<td>.5847</td>
<td>.3143</td>
<td>.5221</td>
<td>.4533</td>
<td>.2721</td>
</tr>
<tr>
<td>III</td>
<td>.5723</td>
<td>.5652</td>
<td>.5847</td>
<td>.5103</td>
<td>.5552</td>
<td>.4533</td>
<td>.6032</td>
</tr>
</tbody>
</table>

At first submission (I), Team 4 and 6 had either wrong format or record count, so it was not able to evaluate their results. The instructor gave feedback to the teams based on their results and discussed with each team about possible ways to improve results. At second submission (II), most teams have improvements on their results and some are significant, e.g. Team 3 from 0.2721 to 0.5847. Before the final submission at the end of the semester, students got chance to enhance their results by asking for evaluation of their intermediate results as input to analyze their process and discussing the results with the instructor back and forth. Finally (III), Team 7 achieved the best results with an index of 0.6032.

For the team project, each team is graded on the design of strategy as well as its final actual results. The purpose of the project is to give students hands-on experience in solving realistic IQ problem using the knowledge and skills gained in the course.

Conclusion

Through its offering and development over the past five years, the IQ tools course has proved to be an important component of the MSIQ curriculum for giving students practical hands-on experience with commonly used IQ tools. Through the data challenge project, they also have the opportunity to design and implement data quality solutions in a team environment.

Although the IQ tools course undergoes continual modification in response to feedback from students and industry stakeholder and the need to remain current with technology, there still many opportunities for improvement. Some of the current challenges revolve around giving remote students the same experience and level of engagement as local students. For example, local students access the software tools on a local server via Remote Desktop. Because of security concerns, student access from off-campus must be through a VPN (Virtual Private Network) connection. Technical issues and the administrative process for VPN approval often cause difficulties and delays for remote students. Due to time zone differences, some students must access the system during their working hours. Even when student’s employer approves class participation during working hours, firewall and security issues may prevent the use of VPN for these students while at work. Because these and other issues, the trend has been to focus more of the class exercises on open source tools such as R, Data Cleaner, and OYSTER that students can download and access with their own systems.

Another problem with the IQ Tools and many other courses in the program is the lack of textbooks. With only a few notable exceptions such as *Journey to Data Quality* (Lee et al. 2006), *Introduction to
Information Quality (Fisher et al. 2008), and Entity Resolution and Information Quality (Talburt 2011), most of the books available for use in the program are written by data quality practitioners for other practitioners. Although there are many advantages in using these books in terms of outlining specific procedures and methodologies, they often tend to be too narrowly focused lacking a general discussion and purposely excluding alternative (competitive) approaches. They also generally lack references to underpin claims and statements, and typically do not provide review questions and planned exercises with test data to reinforce the concepts discussed.

At end, anyone wanting more details on the course content including the tool exercises and challenge project are invited to contact the authors.

References

Lindsey, E. 2008. Three-Dimensional Analysis, Data Profiling LLC.