A Realistic Data Cleansing and Preparation Project

Kwok-Bun Yue
Department of Computer Information Systems
University of Houston-Clear Lake
Houston, Texas, USA
yue@uhcl.edu

ABSTRACT

Although data cleansing and preparation are significant tasks in many real-world data projects, they are rarely found in project assignments in IS database courses. This paper describes a pilot study of a relatively open-ended project assignment in a graduate database course. The project required the students to cleanse and prepare five datasets on educational statistics from United Nations Data before storing them in relations that they designed. To gauge the level of students' prior knowledge on data preparation, the instructor deliberately provided no prior lecture on the topic. A follow-up assignment was a PHP/MySQL Web database application to display educational statistics for a user-specified country. Submitted works and post assignment surveys were studied and analyzed. The result indicated that both assignments were well received and generally beneficial. Although our students appeared not to be well trained in data preparation in their undergraduate studies, they were able to learn quickly enough to produce acceptable products. This approach also appeared to encourage more creativity and better diversity in students' database designs. Our experience suggested that while it was not difficult to identify interesting real-world datasets of appropriate complexity, the instructors will need to put in extra effort on project evaluation. We believe that this kind of assignment can be adapted in many ways to satisfy different educational objectives and it fits well in a well-rounded IS curriculum. Thus, the goal of the paper is to foster interests in real-world data cleansing projects in database courses with a well-examined case study.

Keywords: Database design and development, data cleansing, data management, data modeling, project assignment, Web design and development

1. INTRODUCTION

It is common in database courses to use cleansed data stored in well-defined and simplified relations on common domains for lectures and assignments on query languages including SQL. Examples of these well analyzed and abridged 'toy' domains include employee/department/project and student/course/enrollment (Wagner et al., 2003). Simplified relations of common domains are also employed regularly in popular database textbooks (Hoffer et al., 2008; Elmasri et al., 2010).

For database modeling, students are frequently asked to model a simplified application and design relation schemas accordingly. These relations are then populated with data prepared by either the instructors or the students. For example, in a Web-based multi-media database project, the instructor provided rather specific instructions for designing and populating the database (Holliday et al., 2009). On the other end, in another project, students were required to propose and model their own fictitious applications and populate the designed database in Oracle with their own data (Tuttle, 2002). In either case, the data tend to be relatively clean, well-defined, structured, small in size, single sourced, artificial and simplified.

By contrast, data sources in real-world applications can be dirty, ambiguous, poorly structured, complicated and voluminous (Zhang et al., 2003). Increasingly, companies use a diverse collection of data sources to support their wide range of old and new applications. Unlike traditional clean and internal data, these data sources may be created and provided by external entities. They may not be designed to target a given application of a specific company. In fact, with the advances of Web services, the data producers may not know who may use the data and in what ways they are used. They may design the data format to be generic enough to accommodate the basic shared needs of a large set of clients. Thus, for a given data consumer, before storing in its own database, the data may need to be cleansed, disambiguated, filtered and formatted in order to satisfy application requirements and formats.

To highlight the importance of data cleansing and elaborate existing approaches for improving data quality, Hellerstein (2008) identified four sources of error in databases: data entry, measurement, data distillation and data integration. These errors occur frequently. As a result, significant amount of work is usually spent on data preparation for many data-centric projects. In a survey of 187 data mining projects, 64% indicated that they spent more than 60% of their time on data preparation and cleaning (KDSurvey, 2003). Zhang, Zhang and Yang indicated that "in practice, it has been generally found that data cleaning and preparation takes approximately 80% of the total data engineering effort" (Zhang et al., 2003, pp.375). Likewise, data cleansing is mentioned in MSIS 2000 and MSIS 2006,

the recent model curricula and guidelines for graduate degree programs in information systems (Gorgone et al., 2000; Gorgone et al., 2006). It is also an integral part of the well-known Extraction, Transformation, and Loading (ETL) process that is frequently covered in data warehousing courses (Rahm and Do, 2000).

Using highly simplified, clean and well-defined common domain 'toy' datasets for database courses has the advantages of being easy to teach, use and learn. They provide much value to the student learning process (Wagner et al., 2003) and are thus the main staples of popular database textbooks (for example, Hoffer et al., 2008; Elmasri et al., 2010) and project assignments in database courses. However, they do not prepare IS students to deal with the complexity of real-world data very well. As a result, additional supplementary materials and project assignments on cleansing and preparing realistic data sources can be highly beneficial and effective.

Yet, data cleansing and preparation have neither been well discussed in database textbooks nor well reported in technical papers on database education. Even when they are covered in database textbooks, they are usually discussed very briefly as a part of the ETL process of data warehousing (for example, Kroenke and Auer, 2011; Mannino, 2008). There are a few exceptions in technical papers. For example, in an internship tracking application, students were asked to analyze and cleanse data from 21 Microsoft spreadsheets (DeLorenzo et al., 2011, pp. 375). The cleansing task was relatively simple and focused only on consistent field value descriptions. Isken (2003) created a set of teaching materials and assignments on data cleansing for a decision support course. The target was again placed on specific cleansing and analytic operations on Excel spreadsheets for data analysis and decision support. It was not directly related to databases and no prior programming experience was assumed on the students. Boyno (2003) reported on the experiences on teaching a unit on ETL as a part of a data warehousing course. Students in the course might or might not have DBMS background. The focus was on basic cleansing operations to eventually load the data to a predefined classic star warehouse. Two of the three datasets used were meticulously reconstructed to control appropriate cleansing actions. In general, there seems to be a lack of papers on data cleansing projects in database courses, especially when database design and programming are involved.

To gauge how well data cleaning and preparation is covered in undergraduate database courses, we trawled the websites of all 38 universities with an Information Systems program accredited by ABET (2012) under the information systems curriculum guideline. The focus was on finding courses that include a relatively complete set of course description, syllabus, lecture topics and notes, and homework assignments. Eleven of such courses were identified, with each of them from a distinct university. None of these courses mentioned data cleaning and preparation explicitly in their course description or syllabi, and none included a homework assignment on the subject. We were only able to find one instance of a Powerpoint lecture note on data cleaning. Thus, it appears that a large number of database courses simply do not include data preparation in their syllabi.

Despite their relative importance, there are reasons why data cleansing and preparation are frequently overlooked. Datasets requiring cleansing are usually relatively dirty and messy, just as they happen in the real world. Furthermore, the most suitable techniques are highly project dependent and are thus difficult to generalize. Data cleansing and preparation tools, such as Data Cleaner (2011) and WinPure (2011), can provide powerful features for facilitating data validation and transformation. However, the focus of these features may not satisfy the needs of a specific project effectively. For example, the impressive data type validation tools provided by Data Cleaner are not very helpful for projects where discovering hidden redundancy is important. There is also a substantial learning curve before one can effectively use them. Moreover, explaining the nuances of real world datasets will usually take up much space and effort. Therefore, despite their usefulness, the aforementioned reasons explain their limited coverage in database textbooks and project assignments.

The goal of the paper is to foster interest on real-world data cleansing projects in database courses with a well-examined case study. It aims to help filling the gap in literature coverage by describing an experiment on using a data cleansing and preparation assignment for a graduate database course with realistic United Nations datasets. The remainder of this paper is organized in the following manner. Section 2 describes the goal, design and requirement of the data cleansing and preparation assignment and follows up with a discussion on the kind of data preparation actions that may be conducted. Section 3 analyzes the assignment results to gauge students' prior knowledge and performance in data preparation and database design. Section 4 analyzes the result of a post assignment survey to study its effectiveness from the student perspectives. Section 5 discusses the results of our study and shares the lessons we learnt. We draw our conclusions in Section 6. Appendix 1 lists the assignment.

2. A DATA PREPARATION PROJECT

2.1 The Project Assignment

In our university, graduate Computer Information Systems (CIS) students are required to take a core graduate database course alongside with computer science students. Prior to entering the CIS program, these students are expected to have successfully completed a database course and an introduction to statistics course in their undergraduate studies. The vast majority of our CIS students took these courses from other universities. Our graduate course includes typical topics such as systems analysis for database, data modeling using the Extended Entity-Relationship (EER) model and Unified Modeling Language (UML), the relational model, query languages, SQL, normalization theory, transaction management, etc. Special emphasis is placed on the analysis and design phases on realistic projects to prepare students to adapt quickly to the real world upon their graduation.

In 2011, we decided to add a data preparation and cleansing project to the course. The learning objective was for students to practice retrieving, cleansing and preparing data from a selected data source, and then storing the results in database relations of their own design. Several criteria

were crucial in selecting a suitable data source for the project:
C1. It should be reliable and accessible to ensure data availability during the assignment.
C2. It should be realistic and relatively well-structured to support a reasonably small, interesting and well-defined project.
C3. There should be non-trivial data preparation and cleansing activities that are nevertheless not overwhelming.
C4. It should be interesting and useful enough to support follow-up assignments using the cleansed and stored data.

As a result, a project was assigned in the fall semester of 2011 using United Nations data on educational statistics of countries around the world. An informal experiment was conducted on the assignment with two goals. The first was to launch a pilot study using the assignment as a case study on how well-versed incoming graduate computing students are prepared in cleansing and preparing real-world data. These students should have already taken at least one undergraduate database courses and an introductory statistics course.

Database is a large knowledge area. Many topics can be included in a database course. There is a 'tight time budget' for lectures and assignments. If it is found that the students were already well versed on the subject, lectures and assignments may be skipped to make room for other topics. Also, if students can catch up quickly enough to complete the assignment successfully on their own, an alternative approach will be to provide a brief lecture without an assignment. To accomplish this goal of measuring their prior knowledge on the topic, we did not provide any lecture on data cleansing and preparation before the assignment. Furthermore, the project was structured in a relatively open-ended manner with minimal explanation or instruction before the assignment.

The second goal was to gauge how the assignment may help students to become more proficient in data preparation on real-world domains. To this end, we conducted and analyzed post assignment surveys.

Appendix 1 shows the essential specification of the assignment. For the sake of brevity, the instructions and format requirements on homework submission and storage on a MySQL database are removed. It was designed as an individual project to be completed in two weeks. Basically, students downloaded, analyzed, processed and stored five datasets from the UN Data website: http://data.un.org/Explorer.aspx?d=SOWC (UN Data Explorer, 2011).

As shown in Figure 1 below, the five datasets make up the educational datasets from the Global Indicator Database of the United Nations Statistics Division provided by the site.

The primary education enrollment information shown in Figure 2 provides an example of the nature of data provided by the UN Data website.

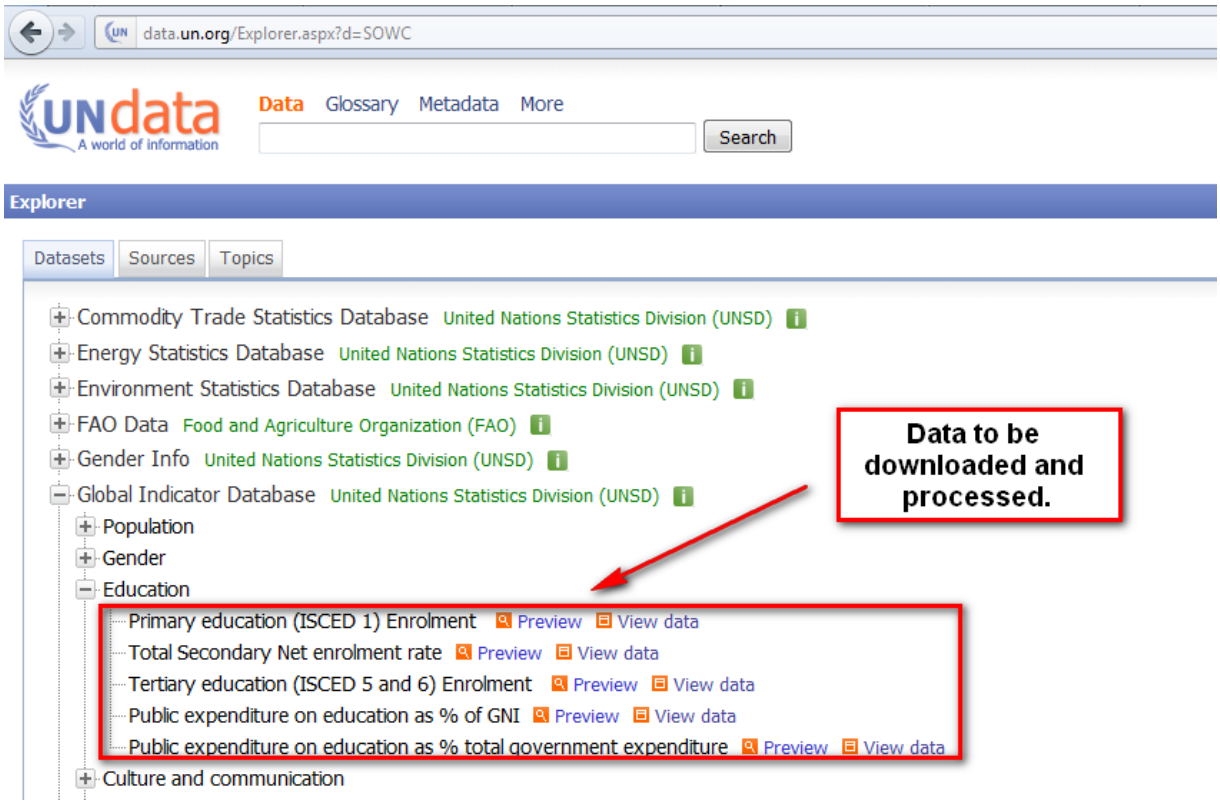

Figure 1. The UN Data website used by the project

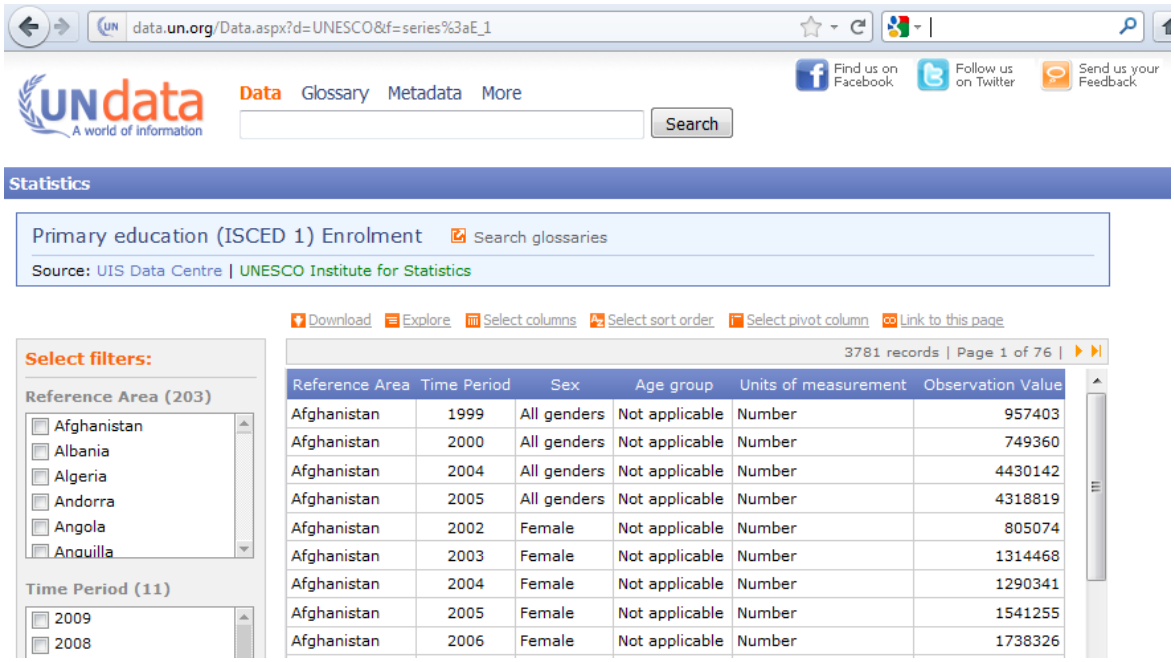

Figure 2. A snapshot of primary education enrollment data provided by the UN Data site

To encourage creativity, the project was relatively open-ended. The instructor provided only a very brief introduction on the importance of data preparation and cleansing with a few generic hints on what data preparation actions students may look for in the assignment. The idea was to stimulate students to explore and discover properties of the datasets independently, to devise data preparation plans, to use the analysis result to design suitable relation schemas, and to eventually store the data in a MySQL database. Students were required to turn in a report answering questions on the cleansing mechanism, tools and actions, the relation design and schemas, and the rationale behind them (Appendix 1).

The five datasets contain student enrollment and educational expenditure statistics for countries around the world for selected years:

DS1. Primary education: contains the numbers of female and/or total primary school students for selected countries and years.

DS2. Secondary education: contains the percentage enrollment rates of female, male and/or total secondary school students for selected countries and years.

DS3. Tertiary education: contains the enrollment number of female and/or total students in tertiary education for selected countries and years. It also contains the female student enrollment as a percentage of total students.

DS4. Public expenditure of education as a percentage of GNI (Gross National Income).

DS5. Public expenditure of education as a percentage of total government expenditure.

Several considerations prompted the selection of these datasets as they seem to satisfy our stated criteria C1 to C4 well.

1. They are useful, interesting and familiar to the students.

2. UN Data is a reputable data source.

3. The size of the datasets is appropriate, with a total number of 13,560 raw records.

4. The datasets are relatively clean, but they still provide plenty of opportunities for cleansing and simplification.

5. The datasets can easily be used for realistic follow-up assignments.

2.2 Data Cleansing and Preparation Activities

The datasets provide a rich collection of opportunities for effective data preparation. We summarize the main ones here. All five datasets have the same flat structure with six identical columns: "Reference Area" (i.e. Country), "Time Period", "Sex", "Age group", "Units of measurement", and "Observation Value". Some columns may *not be applicable* for a specific dataset. For example, the value of the column “Age Group” is always “Not applicable” for all datasets. In another example, the value of the column “Sex” of the two datasets on education expenditures (DS4 and DS5) is always “Not applicable”.

There are also columns within a dataset where their *values* are always the *same*. For example, the value of “Unit of Measurement” is always “Number” for the primary education dataset (DS1) and it is always “Percent” for the tertiary and education expenditure datasets (DS3, DS4 and DS5). These columns can thus be removed without loss of information for these datasets.

There are also *derived records* where a record can be derived from other records within the same dataset. In the tertiary education dataset (DS3), the percentage of female students can be derived from the actual numbers of female students and total students. For example, in the three selected records shown for Afghanistan in 2003 in Table 1 below, the observation value of record #3 (percentage of female students: 20.38839%) can be obtained by dividing the

Rec #	Reference Area	Time Period	Sex	Age Group	Units of Measurement	Observation Value
1	Afghanistan	2003	All genders	Not applicable	Number	26211
2	Afghanistan	2003	Female	Not applicable	Number	5344
3	Afghanistan	2003	Female	Not applicable	Percent	20.38839

Table 1. Three records on tertiary education enrollment (DS3) in Afghanistan in 2003

Reference Area	Time Period	Sex	Age Group	Units of Measurement	Observation Value
Argentina	2003	Female	Not applicable	Number	1253533.737

Table 2. A record with incorrect data value

Rec #	Reference Area	Time Period	Sex	Age Group	Units of Measurement	Observation Value
1	Afghanistan	1999	Female	Not applicable	Number	64110
2	Afghanistan	2000	Female	Not applicable	Number	0
3	Afghanistan	2001	Female	Not applicable	Number	0

Table 3. Primary education enrollment (DS1) in Afghanistan shows missing information

observation value of record #2 (number of female students: 5,344) by that of record #1 (number of total students: 26,211). A simple script and the use of a data cleansing tool both confirmed that this is always the case for all records on female student percentages. Thus, these records can be removed without loss of information. As a result, 1,412 records can be removed in this manner. Furthermore, the remaining records will now always have a value of "Number" for the "Units of Measurement" column and the column can also be removed.

There are also several instances of *incorrect data type values* in the datasets. For example, the number of female students of Argentina enrolled in tertiary education in 2003 is recorded as 1253533.737 in DS3 (Table 2). It should instead be an integer. This is an example of a data type error frequently encountered in data cleansing.

Several actions can be pursued, including contacting the personnel of the data source for confirmation, removing the record, or rounding the observation value. Rounding was deemed to be acceptable after checking against interpolations from the comparable data of nearby years for relative soundness.

An observed value of 0 is also a problem in the datasets. For example, '0' appears in 19 records in the primary education dataset (DS1). As shown in Table 3, it is unlikely that the enrollment literally dropped from 64,110 in 1999 in Afghanistan to 0 in 2000 and 2001. Instead, the zero values are likely meant to represent missing information: *null values*. Thus, these records do not provide any useful information and can be removed.

Besides cleansing, there are also opportunities for *data value simplification*. For example, the values of "Sex" can only be "All genders", "Female" or "Male", requiring 11 characters for storage. They can be simplified to "A", "F" or "M" respectively, thus using only one character.

The UN Data site provides dataset download support in various formats, including XML and comma separated values (CSV). Table 4 summarizes the characteristics of the raw datasets and the resulting datasets after cleansing and simplification. The file sizes are the Byte sizes of the CSV files, which are text files. Thus, 2 to 3 columns could be removed and the file sizes reduced by up to 69.4% of the original raw dataset. Overall, the total number of records of the cleansed datasets is 11,812, a 12.9% reduction from the original 13,560 records in the raw datasets.

Dataset	Raw			Cleaned & Simplified			
	# records	# columns	File size	# records	# columns	File size	% reduced
DS1	3,781	6	256,932	3,765	4	128,930	49.2%
DS2	3,387	6	235,517	3,351	4	122,513	48.0%
DS3	4,319	6	289,427	2,633	4	88,467	69.4%
DS4	1,141	6	85,556	1,140	3	35,284	58.8%
DS5	932	6	70,887	931	3	29,811	58.0%

Table 4. Basic statistics of raw and cleaned datasets

#	Preparation Action	Example	% Student
P1	Remove inapplicable columns	The column "Age Group" in all datasets always has "Not Applicable" as its value.	100%
P2	Remove columns with identical values	The value of the column "Unit of Measurements" in the primary education (DS1) is always "Number".	66.7%
P3	Remove derived records	The records on female student percentages in the tertiary education dataset (DS3) can be derived.	80%
P4	Correctly identify and handle incorrect data type values	A decimal point number appears as the number of students in a very small number of records.	56.7%
P5	Correctly identify and handle null values	A value of 0 appears as the number of students in some records.	33.3%

Table 5. Percentage of student successfully identified and completed data cleaning and preparation actions for every applicable instance

Main tool used	Number of students
Text editor	14
Microsoft Excel	7
Microsoft SQL Server management studio	5
SQL (within MySQL)	4

Table 6. The main data cleaning and preparation tools reported

3. RESULTS OF THE ASSIGNMENT

To gauge how well students explored and cleansed the datasets, possible data preparation actions for the datasets were identified. The submitted assignments of the 30 students in the class were then graded and analyzed to see how well they performed in these data preparation actions. The result is listed in Table 5 above. The last column "% students" indicates the percentage of students that successfully performed the action *for every applicable instance*.

Given that very little background information had been provided before the assignments, the student performance was relatively acceptable. Although nearly all students seem to be aware that the column "Unit of Measurement" has the same value in many datasets, only 20 students removed these columns with identical value completely (action P2). This is because several students selected an approach to use a single master relation to store all five datasets in name-value pair format. After combining the five datasets into a master table, the column "Unit of Measurements" would no longer always have the same value and the students must keep the column. However, this database design approach resulted in relatively poor performance. Successes in actions P4 and P5 were relatively low. Although nearly all students correctly identified some null values (P4) and incorrect data values (P5), many students either did not identify all occurrences or have resolved them unsatisfactorily.

We also monitored some mistaken actions based on failure in data exploration or misconception in the meaning of data. Ten students (33.3%) failed to thoroughly analyze data types that resulted in errors. For example, they might not declare enough space to store the country names which may have up to 52 characters (e.g. "United Kingdom of Great Britain and Northern Ireland"). This resulted in inappropriate truncations. On the flip side, 22 students (71.3%) declared columns that were longer than necessary, resulted in storage inefficiency. A key example of misconception of data semantic is removing the male net enrollment rate column from secondary education (DS2). Six students (20%) mistook that the male net enrollment rate can be derived by subtracting the female net enrollment rate from the total net enrollment rate. Instead, the male net enrollment rate is the ratio of number of enrolled male students to the total number of male population of the right age. It cannot be derived from other data and should not be removed.

We further gauged tools used by the students for data exploration and preparation. Many students used more than one tool. Table 6 lists the *main* tool used by the 30 students.

Most students used a capable text editor as the main tool. The most commonly used was Notepad++ (2011) as it was also used in the class. This is hardly surprising as a text editor is sufficient for the assignment, while being very flexible at the same time. However, only three students reported using regular expressions in Notepad++ to speed up their work, indicating that their proficiencies in using text editors may not be in the expert level. Since all students started with CSV, it is also natural that Excel was the second most popular choices.

The use of Microsoft's SQL Server management studio by five students was a little surprising. The studio provides tools for data import, analytics, integration and management that can be useful for the assignment (MS SQL Server, 2011). However, since the requirement was to store the tables in MySQL, students using the tool would need to import the raw data into MS SQL Server, clean and merge the data within MS SQL Server, export the tables, and finally import them into MySQL. The overall quality of the works submitted by these five students was below the class average. This is mostly due to the fact that the employed SQL Server-centric approach did not match well with the assignment. It seems that the students simply selected the tool that they were most comfortable with, ignoring suitability. This is an example showing why students need to

broaden their 'tool boxes' and be able to analyze and select the right tool for a specific problem. One tool does not fit all.

Lastly, four students imported the raw data into MySQL and used SQL statements to explore, clean and transform the tables. Their works were actually better than the class average and these students show better mastery of SQL. However, it can also be seen that for the same preparation actions, their solutions were significantly more complicated and demanding than those students using text editors or Excel. In this project, it is not as easy to use SQL statements to perform the needed cleansing and preparation. This again indicates the importance of picking the right tool.

To supplement their main tools, only four students have reported using dedicated tools developed specifically for data exploration, cleansing, and pre-processing. These tools can be powerful. For example, Data Cleaner (2011), an open source tool used by most of these students, has powerful features for "analyzing, profiling, transforming and cleansing data." However, students in general were not familiar with them and there was a steep learning curve. Most of their features are also not needed for the assignment. Given the shortness of the assignment, most students simply did not invest much time in identifying and learning these tools.

Overall, our experience indicated that our student background on data cleaning and preparation might be less than desirable. Their performance was further impacted by the deliberate lack of coverage of the topic before the assignment and the tight schedule. On the other hand, it seems that the students can learn quickly enough to produce works that are in general acceptable. Future assignments with proper lectures and longer schedules should produce even better results. Alternatively, an instructor may select to provide a brief lecture without an assignment, with the understanding that the students can quickly learn enough to perform adequately if needs arise. In fact, this is the approach we used in the next semester, while employing this project as a case study for the students. The data cleansing assignment was replaced by an XML DB assignment.

For database design, there appear to be very few obvious mistakes besides that some columns might sometimes be declared with inappropriate length. Student considerations can be classified into several groups below. Note that the sum of the percentages of students adopting the considerations is greater than 100% as a student may use more than one consideration.

D1. Direct mapping (employed by 70% of students). Students might map one dataset to a single relation with inapplicable columns removed. This is the most straightforward approach and it resulted in five relations, one per each dataset.

D2. Generic master table (23.3%). Students might store all datasets in a master table that basically include three columns to store the name, data type and value of any educational parameter of a country in a given year respectively. The advantage is its extensibility to easily incorporate new and similar datasets. However, because of the generality, the SQL statements for extracting necessary information become more complicated and significantly less efficient, making it less appropriate for the assignment.

D3. Additional tables for allowed values (33%). Students might also create additional tables to store allowable values. For example, a country table may be created to store short country codes and the corresponding names. This approach enhances data integrity at the cost of a possible slight decrease in performance.

Although we have not done any formal analysis, our experience indicated that there was more diversity in student database designs when compared to our previous database design assignments using simplified toy applications. We posted grading notes about these varieties of designs together with a suggested solution to the students to promote learning by comparisons.

4. SURVEY RESULTS

To gauge the student perception of the assignment, an anonymous survey was conducted after the solution was posted. On a scale of five, the survey simply asked the students how useful, interesting and difficult the assignment was. They also might provide written comments. Table 7 shows the responses from 25 students.

It should be noted that the survey result is not scientific because of the size and sample limitations. Furthermore, since surveys have not been conducted on most of the other assignments, there is a lack of reference points for the comparative effectiveness of the project. Nevertheless, the result provides some initial insight as a pilot case study. In general, the survey suggests that the students found the assignment to be both useful and interesting, but more useful than interesting. The students also found the assignment to be a little difficult. This reinforces our initial perception of the students' relative lack of their background in data cleaning and preparation.

Twenty four students provided written comments in the survey, some of which were long. We classified the comments into general groups of opinion. Table 8 shows the number of student responses for each general opinion. A student feedback may include more than one opinion.

Overall, the survey may be considered as a good data point to support the perceived effectiveness of this kind of approach used for designing the assignment. The students especially liked the real-world nature and the exposure to data cleaning and preparation. Some expressed that the project is hard. To make the assignment less difficult for the students, logistical arrangements can be improved to ease the burden on the students. To gauge the student background in data preparation, by intention, minimal lecturing and background information of the assignment had been provided. This can easily be changed in the future and suitable lectures prior to the assignment should greatly alleviate the situation. In fact, that was the main suggestion for improvement provided by the students in the survey.

How did the students found the assignment?	Average response
Very uninteresting (1) to very interesting (5)	4.12
Very unuseful (1) to very useful (5)	4.42
Very easy (1) to very difficult (5)	3.69

Table 7. Average student response of assignment survey

Opinion	Number of students (percentage)
Like the real-world nature of the project	16 (67%)
Express that the project provided good exposure to the importance of dataset selection, data preparation and cleansing	13 (54%)
Express that the project is useful and/or good	10 (42%)
Express that the project is interesting and/or challenging	5 (21%)
Think that the project is hard or relatively hard	5 (21%)
Express in a positive way that the project is very different than assignments they have done before	4 (17%)
Provide suggestions for improvement	3 (12.5%)

Table 8. Opinions in the written comment section of the survey

5. DISCUSSION

Our experience suggests that it is not difficult to identify good real-world datasets that satisfy our selection criteria C1 to C4. Once the real-world datasets are identified, one of the disadvantages of working with them is that the instructor needs to work hard to identify suitable cleansing and preparation actions. However, finding interesting and useful datasets to satisfy a specific course objective should not be difficult.

The complexity of the assignment can easily be adjusted to fit course and project objectives. Our assignment was limited in scope and depth because it accounted for only 12.5% of the grade for all assignments. However, it can be adapted and extended in many directions. For example, the five datasets are stored in five relations in MySQL. These five datasets can be merged into a single master relation using the columns Country and Year as a composite primary key. Figure 3 shows the SQL statement in MySQL to create such a master relation in which all information for a given country and year is stored in a single record. The meaning of each column should be obvious.

This approach results in reduced storage and faster SQL execution. It was instructive to the students to learn that the merging of data can be done in various ways:

M1. Merge the five datasets using text editors or data preparation tools into a single dataset and store it in the database using a load statement.

M2. Develop a script program to read from the five cleaned datasets, merge the data, and store them into the database.

M3. Use SQL statements to merge the five relations into a single master relation.

This variety of approaches also reinforces the open-ended nature of data cleansing and preparation and highlights the fact that there are usually many ways to accomplish the same task. Although the assignment did not ask the students to merge datasets, we developed a standalone PHP program for method M2 and showed it to the students after the assignment. Note that PHP is usually used as a Web server-side scripting language and not used in standalone mode. We selected it simply because PHP would be used in the next assignment and thus the script program could serve as a valuable example. Method M2 can be implemented easily in other scripting languages such as VBScript, Perl, Python or Ruby. The feedback on the script program was very good as it further expanded and consolidated what the students have learnt from the project.

A major advantage of this assignment is the availability of an interesting and clean database with proper relation schemas. This facilitates follow-up assignments to provide continuity and depth. For example, it is easy to construct various subsequent data mining or data warehousing assignments. In our case, the next assignment was a PHP MySQL Web application using the master relation created by method M2 to display all the available educational statistics for a user-specified country. Again, because it is only one out of eight assignments, the requirement was minimalistic, with the emphasis being on the Web database connectivity using PHP, and not on the design and user interface of the Web pages. In the assignment, the user should be able to use a simple interface to specify the desired country, such as the one shown in Figure 4.

```
create table UN_education (
  country varchar(52) not null,
  year integer(4) not null,
  female_primary integer(10) unsigned,
  all_primary integer(10) unsigned,
  female_secondary decimal(8,5) unsigned,
  male_secondary decimal(8,5) unsigned,
  all_secondary decimal(8,5) unsigned,
  female_tertiary integer(10) unsigned,
  all_tertiary integer(10) unsigned,
  gni_percent decimal(8,5) unsigned,
  exp_percent decimal(8,5) unsigned,
  unique key (country, year)
);
```

Figure 3. The MySQL statement for creating a single master relation for the datasets

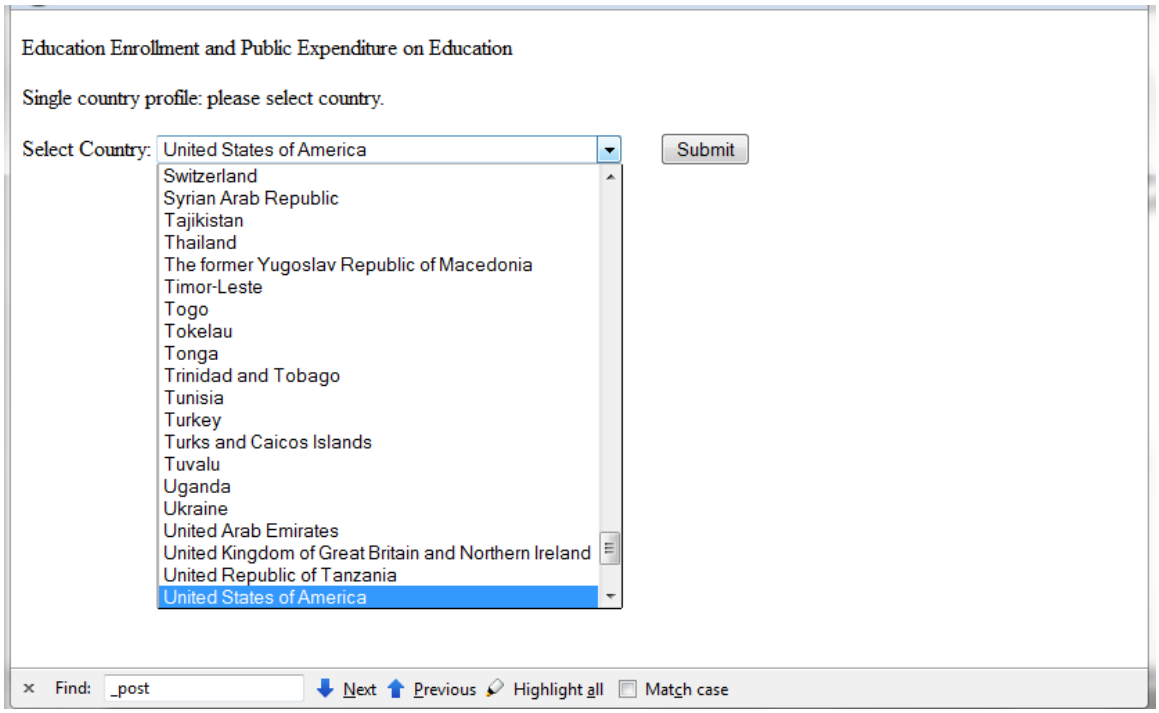

Figure 4. A simple interface of a PHP Web database application for showing educational statistics

Education enrollment and public expenditure statistics:

Country: United States of America

Range of years with data: 1999 to 2008

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Number of female primary students	12337903	12085114	12329668	12181565	12139802	11807911	11871914	11923213	11945630	12071989
Number of all primary students	24937931	24973176	25297600	24855480	24848518	24559494	24454602	24319033	24492041	24676574
Enrollment rate of female secondary students		86.9%	87.4%	84.7%	88.5%	89.8%	89.6%	89.1%	89.1%	89.0%
Enrollment rate of male secondary students		84.9%	85.4%	84.6%	87.2%	87.4%	87.2%	87.4%	87.4%	87.7%
Enrollment rate of all secondary students	87.5%	85.9%	86.4%	84.7%	87.9%	88.6%	88.4%	88.2%	88.2%	88.3%
Number of female tertiary students	7663037	7362121	7596436	8967172	9409595	9644920	9884782	10031550	10184055	10432214
Number of all tertiary students	13769362	13202880	13595580	15927987	16611711	16900471	17272044	17487475	17758870	18248124
Public education expenditure as a % of GNI	5.0%		5.7%	5.6%	5.8%	5.6%	5.2%	5.6%	5.4%	
Public education expenditure as a % of government expenditure			17.1%	15.2%	15.2%	14.4%	13.7%	14.7%	14.1%	

Figure 5. The result of the query on USA of Figure 4

The Web application should then show all educational statistics for the selected country (Figure 5).

We also conducted a similar anonymous post assignment survey. The result is shown in Table 9.

The survey result indicates that the students received the follow-up assignment even better and they found it to be more interesting and useful. Since 60% of respondents reported no prior programming experience in PHP, they also found it more difficult. By reading the written comments and discussion with the students, we found that the students enjoyed well specified and 'concrete' application development. This resulted in better reception of the assignment. The students also seem to have less experience and feel less comfortable on open-ended projects, such as our data cleansing and preparation assignment. This actually speaks for the need of more open-ended assignments in IS education as the real world has both kinds of projects: well specified or relatively open-ended.

How did the students found the assignment?	Average response
Very uninteresting (1) to very interesting (5)	4.33
Very unuseful (1) to very useful (5)	4.53
Very easy (1) to very difficult (5)	3.93

Table 9. Average student responses of the survey of the follow-up assignment

There are many other possible ways to adapt the assignment to satisfy different course objectives. Hellerstein (2008) classified two types of data for cleaning:

1. Quantitative data: integers or floating point numbers that measure quantity of interest.
2. Categorical data: for specifying data into categories or groups.

He also elaborated two important special cases of categorical data cleaning: postal addresses and identifiers (keys). Different techniques are used for these four cases, such as outlier detection for quantitative data and ontological techniques for categorical data. Although our assignment contains both quantitative and categorical data, the cleansing actions are mainly focused on quantitative data. It is conceivable that data cleansing assignment focusing on categorical data would fit specific database or data mining courses better.

Another worthwhile adaptation is to set the assignment as a small team project on more complex datasets with more time to allow the team to research appropriate cleansing tools. A team approach may be more suitable for this kind of open-ended projects in which team members can complement each other well. Other possible ways of adaptations include using datasets that are more complicated and not flat in structures, using dedicated data cleansing and preparation tools as a requirement, using multiple external data sources, and using both external data sources and internal database information.

Despite its usefulness, there are disadvantages of this kind of open-ended real-world dataset approach. The instructor usually needs to work harder to construct a reasonable project. Grading can also be a problem because of the open-ended nature. Any standard solution may be less universal and more difficult to construct. Creating a repository of case studies and applying innovation in assignment evaluations may help in this respect.

6. CONCLUSIONS

This paper describes a pilot case study on a data cleansing and preparation assignment based on real-world datasets. Such an assignment is lacking in many database courses. The study finds that our graduate students were not well exposed to data preparation in the real world. However, they were able to learn enough to successfully complete the assignment with a reasonable amount of effort. With proper preparation, the assignment can be both effective and beneficial. It can also be adapted in many ways to satisfy different learning objectives in the graduate IS curriculum. Possible extensions of this work include expanding the approach on projects with different kinds of datasets, assignments and courses; using a team project setting; and establishing a library of related case studies.

7. ACKNOWLEDGEMENTS

We would like to thank our students for their interest, participation, and feedback on this experiment. Ms. Chloris Yue and scholars of the UHCL NSF Scholar Program (NSF Grant # 1060039) also provided invaluable suggestions and assistance.

8. REFERENCES

ABET (2012), "Find Accredited Program", Retrieved April 27, 2011, http://main.abet.org/aps/Accredited programsearch.aspx.

Boyno, E.A. (2003) "Extraction, Transformation and Loading in a Data Warehouse Course," Information Systems Education Journal, Vol. 1, No. 10, pp. 3-9.

Data Cleaner (2011), Retrieved December 20, 2011, from http://datacleaner.eobjects.org/.

DeLorenzo, G. J., Kohun, F.G., Nord, D. and Nord, J.H. (2011) "Integrating Service Learning and Civic Engagement Opportunities into Professionally Accredited Business and IS Programs in the US and Europe to Enhance Student Learning Outcomes, Research, and Local Community/Economic Development," Proceedings of Informing Science IT Education Conference, pp.153-162.

Elmasri, R. and Navathe, S. (2010) Fundamentals of Database Systems, 6th Edition, Addison-Wesley, Boston, MA.

Gorgone, J., Gray, P., Feinstein, D. L., Kasper, G. M., Luftman, J. Stohr, E. A., Valacich, J. S., and Wigand, R. (2000) "MSIS 2000: Model Curriculum and Guidelines for Graduate Degree Programs in Information Systems," Communications of the Association for Information Systems, Vol. 3, Article 1. Available at: http://aisel.aisnet.org/cais/vol3/iss1/1.

Gorgone, J. T., Gray, P., Stohr, E., Valacich, J. S., and Wigand, R. T. (2006) "MSIS 2006: Model Curriculum and Guidelines for Graduate Degree Programs in Information Systems," Communications of the Association for Information Systems, Vol. 17, Article 1. Available at: http://aisel.aisnet.org/cais/vol17/iss1/1.

Hellerstein. J. (2008) "Quantitative data cleaning for large databases," United Nations Economic Commission for Europe, Retrieved December 22, 2011, from http://db.cs.berkeley.edu/jmh/papers/cleaning-unece.pdf.

Holliday, M.A. and Wang, J. (2009) "A multimedia database project and the evolution of the database course," Proceedings of the 39th IEEE international conference on Frontiers in education conference (FIE'09), IEEE Press, Piscataway, NJ, pp.1092-1097.

Hoffer, J., Prescott, M., and Toppi, H. (2008) Modern Database Management, 9th Edition, Prentice-Hall, Upper Saddle River, NJ.

Isken, M. (2003) "Data Cleansing and Analysis as a Prelude to Model Based Decision Support," INFORMS Transactions on Education, Vol. 3, No. 3, pp.23-75.

KDSurvey (2003) Retrieved December 20, 2011, from http://www.kdnuggets.com/polls/2003/data_preparation.htm.

Kroenke, D, and Auer, D. (2011), Database Processing, 12th Edition, Prentice-Hall, Upper Saddle River, NJ.

Mannino, M. (2008) Database Design, Application Development, and Administration, Fourth Edition, Ediyu.

Microsoft's SQL Server (2011) Retrieved December 21, 2011, from http://www.microsoft.com/sqlserver/.

Notepad++ (2010) Retrieved December 22, 2011, from http://notepad-plus-plus.org/.

Rahm, E. and Do. H. H. (2000) "Data cleaning: Problems and current approaches," IEEE Data Engineering Bulletin, Vol. 23, No. 4.

Tuttle, S. M. (2002) "Practical lessons from experience with the database design course project," Journal of Computer Science in Colleges, Vol. 18, No. 2, pp.32-42.

UN Data Explorer (2001) Retrieved December 20, 2011, from http://data.un.org/Explorer.aspx?d=SOWC.

Wagner, P., Shoop, E. and Carlis, J. (2003). "Using scientific data to teach a database systems course," Proceedings of the 34th SIGCSE technical symposium on computer science education (SIGCSE '03), New York, NY, pp.224-228.

WinPure (2011) Retrieved December 22, 2011, from http://www.winpure.com/Data-Cleansing-Tool.html.

Zhang, S. Zhang, C. and Yang, Q. (2003) "Data Preparation for Data Mining," Applied Artificial Intelligence, Vol. 17, pp. 375-381.

AUTHOR BIOGRAPHY

Kwok-Bun Yue (B.S., M.Phil., Chinese University of Hong Kong, M.S., Ph.D., University of North Texas) is a Professor of Computer Information Systems and Computer Science at University of Houston-Clear Lake (UHCL). His research interests are in Internet computing, semi-structured data, and information systems and computer science education. He had published more than 30 technical papers. Dr. Yue is a recipient of s UHCL Distinguished Teaching Award, the UHCL Piper Award, the UHCL Alumni Association's Outstanding Professor Award and the UHCL Fellowship Award. He had served as a CTO of a startup company.

APPENDICES

Appendix 1.The Essential Part of the Data Preparation Assignment with Submission Instructions Removed

As the volume and complexity of data increase quickly, data preparation becomes an important issue for database and data mining. This assignment combines data preparation and simple database design and will serve as the template of the next assignment on SQL/PHP.

Eventually, we will build a Web application to show selected statistics of education for different countries in the world. There are many public datasets available. We will use datasets from UN Data:
http://data.un.org/Explorer.aspx?d=SOWC

In "Global Indicator Database", there are five datasets:

- Primary education (ISCED 1)
- Total Secondary Net enrolment rate
- Tertiary education (ISCED 5 and 6) Enrolment
- Public expenditure on education as % of GNI
- Public expenditure on education as % total government expenditure

Your task is to download all five datasets, clean and simplify them, design relational schema and store all information of the data in the database.

The site provides various download options, select the ones that are most suitable for your tasks.

Study the data to clean and simplify it. For example, selected fields may have the same values for all entries and are thus not needed to store in a separate column in a relation. Some field values may be derivable from other fields

.

Turn in:
(1) A report on how you obtain, clean and store the data.
(a) What format of the data have you downloaded and why?
(b) What properties have you found about the five datasets that can be used to prepare and simplify the data?
(c) What mechanisms have you employed to clean and prepare the data and why?
(d) What are the relation schemas of the tables you have designed to store the data? Identify all candidate keys.
(e) What are the SQL commands you have used to create the tables and populate them?
(f) What are the format of the cleaned data files you used to populate the tables, and why?

[For brevity, the instructions and format requirements on homework submission and storage in a MySQL database are removed below]
…