Addressing Student Difficulties in Using Numeric Data Downloaded From the World Wide Web

Marilyn K. Pelosi
School of Business
Western New England College
Springfield, MA 01119

David L. Russell
School of Business
Western New England College
Springfield, MA 01119

Abstract

The authors describe three circumstances under which a student's ability to download data from the World Wide Web can be complicated, and requires a level of skill beyond what can be reasonably expected of casual student users. The assumption is made that the downloaded data is to be subsequently analyzed in commonly-available data analysis tools. They conclude with suggestions to faculty and administrators on ways students can be supported in downloading data.

Keywords: World Wide Web, data download, quantitative analysis

Marilyn K Pelosi, Ph.D., Professor of Quantitative Methods, 413-782-1713,
mailto:mpelosi@wnec.edu?subject=Feedback from ISECON 2003 paper
David L. Russell, Ph.D., Associate Professor of Computer Information Systems, 413-782-1479,
mailto:drussell@wnec.edu?subject=Feedback from ISECON 2003 paper

1. INTRODUCTION

In this paper, we address an increasingly common difficulty for students: how to take numeric data found on the World Wide Web and perform transformations so that commonly-available tools can be used for quantitative analysis. One might react by questioning why this is difficult. Indeed, persons well experienced in the use of Web-generated data might consider the problem trivial. We propose, however, that students have great difficulty in downloading and transforming numeric data from the World Wide Web, and that is particularly true for students in the Arts and Sciences.

All too frequently, professors in a variety of disciplines expect students to find relevant numeric data on the Web, download the data to files and then analyze their content using a common computing tool. Unfortunately, this work is assigned to students on the assumption that the required numeric data are available. Further, if the data are available, the professor assumes that they can be used straightforwardly with a data analysis tool available to students. We will demonstrate that this is often far from the case by presenting three scenarios.

Moreover, many students, particularly those studying in Liberal Arts and the Humanities,

have relatively little instruction in computing skills. If any instruction is provided, it generally consists of a course with three sections: one each for word processing, spreadsheets and databases, corresponding generally to the Microsoft Office™ tools: Word™, Excel™ and Access™. Instruction in a web browser and other Office tools is generally absent. The level, intensity and breadth of focus is far less than the IS 2002.PO course found in the IS 2002 model curriculum (Gorgone, *et al*., 2002) and it can be said that this student is largely self-taught. For the purposes of this paper, we identify this student as a "casual student user".

Thus, this paper focuses on the use of such Web-derived data, with particular attention to the conversion of these data into a form and format in which they can be computed. Further, this paper addresses faculty outside of the IS discipline, as well as those outside of Business and Engineering, that can be reasonably expected to make use of both web search tools and spreadsheets in analysis. We will assume that the data analysis tool is spreadsheet software; in particular, we will use Microsoft Excel™ 2002. This assumption can be justified for two reasons: first, this tool is ubiquitous in colleges and universities; and second, Microsoft Excel's workbook file format (*.xls), can serve as a receiving file format for many other tools, such as SPSS™ and Lotus 1-2-3™.

These assumptions imply a series of hurdles that the student must surmount:

1. The student must find appropriate data using a search engine;
2. The student must download the data in such a manner that they can be converted for analysis; and
3. The student must then use the analysis tool correctly.

In this paper, we will assume that the student has sufficient mastery in steps 1 and 3. Hence, we will concentrate on the second step: the downloading of data from the Web and their transformation to a form and format in which they can be used by a data analysis tool. To do this, we present three different scenarios and describe the challenges students face. More specifically:

1. Some data are in a form that is not conducive to direct and effective use in a data analysis tool;
2. Even data made available in a standard format often cannot be used without editing;
3. Some data are presented in a non-tabular layout that causes their transformation to a standard file format to be very challenging; and
4. Data must frequently be downloaded in a text format which students must then import the text format into the analysis tool (In Microsoft Excel 2002, the means the use of the Text Import Wizard.)

Using Excel as an example, the last step above means the following steps must be done:

1. Determine column breaks that do not necessarily conform to the column breaks in the downloaded data;
2. Make correct decisions about data types in each column;
3. Actually convert the text file into an *.xls file;
4. Do extensive editing in Excel to address inconsistencies; and
5. Remember to save the file in *.xls format rather than the default format, the native *.txt format

Through the use of scenarios, we will achieve the following objectives:

1. to describe situations in which some downloaded data is in a form that is not conducive to effective use;
2. to address situations in which data to be downloaded is not available in any standard file format;
3. to address situations in which data must frequently be downloaded in a

text format, and the consequences of this action;

4. to address converting the text format into a usable file format; and
5. to address the data structures and editing actions required to make the converted data file usable.

Our three scenarios are:

1. Scenario 1: Web sites which make data available in file formats that can be used directly by data analysis tools, the large majority of which are sites making files available in Excel's native file format (*.xls);

2. Scenario 2: Web sites that make data available in text format. Here, we make a distinction between:

   a) sites that offer data in a tabular structure, in text format and are structured in such a way to facilitate its transformation (Scenario 2a) versus

   b) those that offer data in a tabular structures, but with complications that inhibit the direct use of the data without significant editing by the student (Scenario 2b); and

3. Scenario 3: Websites whose data is not in tabular structure, and which require significant work to ultimately perform in a common analysis tool.

2. SCENARIO 1: DATA AVAILABLE IN STANDARD FILE FORMATS

Here we address what most would consider the least challenging scenario: the use of data that that is available in Excel (*.xls) format. Often, however, (1) the data are not in true tabular form; (2) the data contain control breaks in the form of sub-summation and summarization lines which inhibit the student from analyzing the entire block of data and (3) the data require editing chores to make the data usable. An example of this scenario follows.

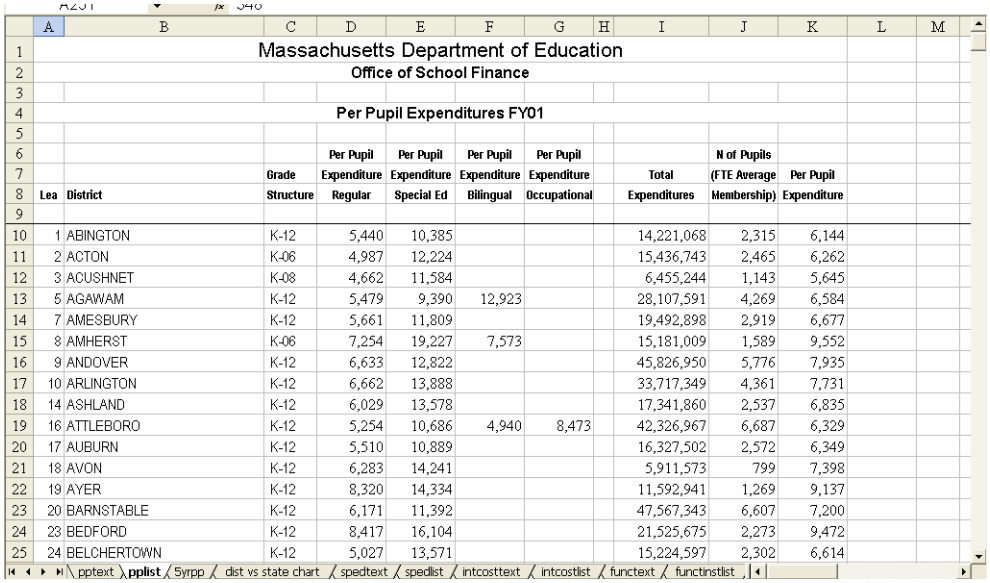
Say that a Education professor has assigned students a project in which they must analyze per-pupil expenditures in the Commonwealth of Massachusetts. In particular, the student must research the relationship of per-pupil expenditures and population in the 351 cities and towns that make up the state. In Massachusetts, a "town" is comparable to a "township", as that term is known in most of the United States. There is no equivalent to a "town" (that is, a six mile square geographical entity) in Massachusetts.

Let us assume that the student has used a browser and has sufficient skills to find

http://finance1.doe.mass.edu/statistics/

which is the web site of the Massachusetts Department of Education. Once the site has loaded, the student will see several downloadable Excel data files, one for each recent fiscal year up to fiscal year ("FY") 2001 (the latest year for which data are available). When the FY 2001 option is selected, an Excel data file is opened within the browser. The workbook, entitled 'Copy of pp01.xls', contains 11 worksheets. Four are textual introductions of the various topics; each introduction is followed by one or more worksheets of data. The data worksheets contain data in highly usable tabular format, an example of which is shown in Figure 1.

Likely, the first data worksheet the student would use is to the far left. This worksheet is labeled 'pplist' and titled 'Per-Pupil Expenditures FY01'. Unfortunately, the data on this worksheet present challenges to the student. For example, the worksheet lists values by local education agency ("LEA"). There are 328 LEAs, which is less than the 351 cities and towns in the Commonwealth. Although in most cases "LEA" and the city or town are synonymous, the difference is due to regional school districts (since Massachusetts has a high population density, rural areas and thus regional school districts are rare). This is significant since all other data from the Commonwealth is organized alphabetically by the 351 cities and towns. The LEAs are listed in rough alphabetical order by town, but with a number of exceptions. The LEA field (column A) also holds code numbers from 1 to 915, with many gaps. Finally, the data contain several layers of summarization. For example, summary data for metropolitan



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Massachusetts Department of Education												
2	Office of School Finance												
3													
4	Per Pupil Expenditures FY01												
5													
6				Per Pupil	Per Pupil	Per Pupil	Per Pupil			N of Pupils			
7			Grade	Expenditure	Expenditure	Expenditure	Expenditure		Total	(FTE Average	Per Pupil		
8	Lea	District	Structure	Regular	Special Ed	Bilingual	Occupational		Expenditures	Membership)	Expenditure		
9													
10	1	ABINGTON	K-12	5,440	10,385				14,221,068	2,315	6,144		
11	2	ACTON	K-06	4,987	12,224				15,436,743	2,465	6,262		
12	3	ACUSHNET	K-08	4,662	11,584				6,455,244	1,143	5,645		
13	5	AGAWAM	K-12	5,479	9,390	12,923			28,107,591	4,269	6,584		
14	7	AMESBURY	K-12	5,661	11,809				19,492,898	2,919	6,677		
15	8	AMHERST	K-06	7,254	19,227	7,573			15,181,009	1,589	9,552		
16	9	ANDOVER	K-12	6,633	12,822				45,826,950	5,776	7,935		
17	10	ARLINGTON	K-12	6,662	13,888				33,717,349	4,361	7,731		
18	14	ASHLAND	K-12	6,029	13,578				17,341,860	2,537	6,835		
19	16	ATTLEBORO	K-12	5,254	10,686	4,940	8,473		42,326,967	6,687	6,329		
20	17	AUBURN	K-12	5,510	10,889				16,327,502	2,572	6,349		
21	18	AVON	K-12	6,283	14,241				5,911,573	799	7,398		
22	19	AYER	K-12	8,320	14,334				11,592,941	1,269	9,137		
23	20	BARNSTABLE	K-12	6,171	11,392				47,567,343	6,607	7,200		
24	23	BEDFORD	K-12	8,417	16,104				21,525,675	2,273	9,472		
25	24	BELCHERTOWN	K-12	5,027	13,571				15,224,597	2,302	6,614		

pptext | pplist | Syrpp | dist vs state chart | spedtext | spedlist | intcosttext | intcostlist | functext | functinstlist

Figure 1: Downloaded FY 01 data, 'pplist' worksheet

areas and counties are listed within the data, although the cities and towns that make up these figures are also listed; this would inhibit the generation of correct summary data. Thus, despite the title of the worksheet, per-pupil expenditure data for all the cities or town are not found on this worksheet. A student using this data inadvertently unwittingly impairs his or her subsequent analysis.

The information the student seeks, per-pupil expenditure for all the 351 cities or towns, is available on another worksheet located six worksheets to the right of the 'pplist' worksheet. This worksheet is labeled 'incostlist' and titled "FY01 Integrated Cost Per Pupil". In fact, "integrated" means the costs of multi-town regional school districts are spread or "integrated" across the constituent towns. However, there is no indication of which school districts (LEAs) are composed of an individual city or town and which are part of a regional district.

In the 'intcostlist' worksheet, the student also faces obstacles. While this worksheet lists the 351 cities and towns of the Commonwealth, the 'LEA' code from the first worksheet is not provided. There are also no categorical data, such as county, region, metropolitan area or SMSA, which would handicap later categorical analysis, such as the use of PivotTables in Excel. Further, the population of the city or town (which is a requirement of the assignment), is not provided, inhibiting the student's assignment to integrate per-pupil expenditures and population.

To integrate population data and categorical data (such as county or SMSA), data will have to be imported from other information sources about the Commonwealth such as MISER, a research organization at the University of Massachusetts. After a small amount of searching of moderate difficulty, the student finds:

http://www1.miser.umass.edu/datacenter/population/MISEREstim1998/miserest.html

in which 1999 estimated population data for the 351 cities and towns (the latest year available) is listed in a downloadable Excel worksheet. An example of the data is shown in Figure 2.

Now the student confronts further challenges. The data are not fully tabular, as they contain frequent summarization lines and frequent blanks lines for readability. Further, unlike the data from the Department of Education, these data are sorted by county and sub-sorted by city or town. The data also contain a plethora of data fields which are not relevant to the investigation of per-pupil expenditures. An experienced

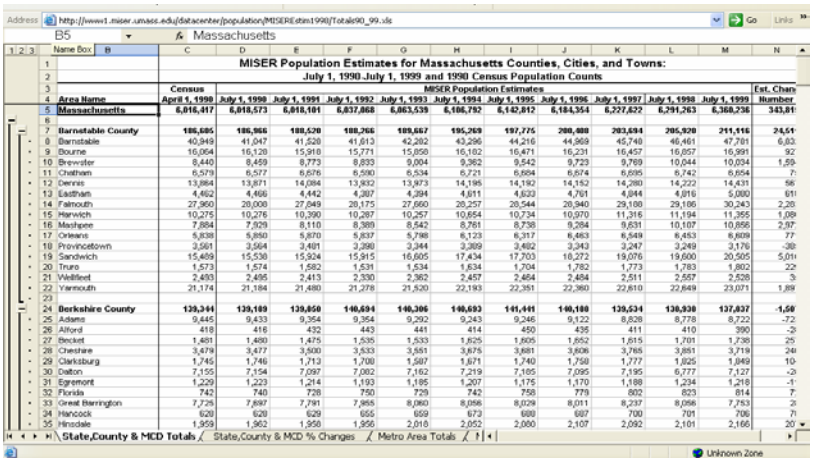


Fig. 2: downloaded MISER data

Excel user would:

1. delete unneeded rows and columns (in particular the intermediate summary rows);
2. make the reference to county a categorical variable;
3. resort the data by city or town; and finally
4. copy the data to the 'intcostlist' worksheet.

However, it is unrealistic to expect such a skill level in a casual student user.

To provide an example of the unrealistic demands on the student, we present below the following steps a student would have to take to make the data usable (This and all subsequent worksheets are available are in an Excel workbook available from the authors upon request):

1. Using the Edit – Move or Copy Sheet function, create a temporary copy of the 'State,County & MCD Totals' worksheets;
2. In this worksheet, manually eliminate all spacing lines, county headers, unneeded columns and summary data rows;
3. Sort the data by city or town, so that they are in the same order as the 'intcostlist' worksheet.
4. Copy these data onto a copy of the 'intcostlist' worksheet.
5. In the resulting merged worksheet, the city or town names from 'intcostlist' are in column B while the city or town names imported from MISER are in column G.
6. However, the values from the Department of Education and those from MISER often do not match. Thus, the student must create a logical test column (here, column J) to determine where there are any mismatches in the data:

```
=IF([value in column B]=
      [value in column G]),
           "OK","Out of sequence")
```

7. This logical test reveals an immediate problem: since the IF function's logical test is case-specific, no values in the original data in column B (which is all in upper-case format, e.g., 'ABINGTON') match the values in column G (which is in title format, e.g., 'Abington'). Therefore the student must modify the logical test in column J to force column G into upper case:

```
=IF([value in column B]=
      UPPER([value in column G]),
           "OK","Out of sequence")
```

(Alternatively, both values could be converted to lower case through the

use of the LOWER function, or to title format using the PROPER function.)

8. Unfortunately, the student soon discovers that even this will not work, because column B's values contain literal blanks at the end of the entered values. For example the first value is actually 'ABINGTON' followed by 21 blanks, that is:

```
ABINGTONxxxxxxxxxxxxxxxxxxxxx.
```

9. This in turn requires a re-write of the logical test in column J to incorporate the length of the city or town name. For example, row 10 would be:

```
=IF(LEFT(B10,LEN(G10))=
        UPPER(G10),
             "OK","Out of sequence")
```

10. This results in 36 records that are "out of sequence". These are dominated by sorting errors, that is, an error of incorrect placement of values under the collating sequence, indicating that one or both of the original tables was sorted incorrectly. An example of sorting errors can be seen in values 85-87 (spreadsheet lines 94-96):

EASTHAM	East Longmeadow
EASTHAMPTON	Eastham
EAST LONGMEADOW	Easthampton

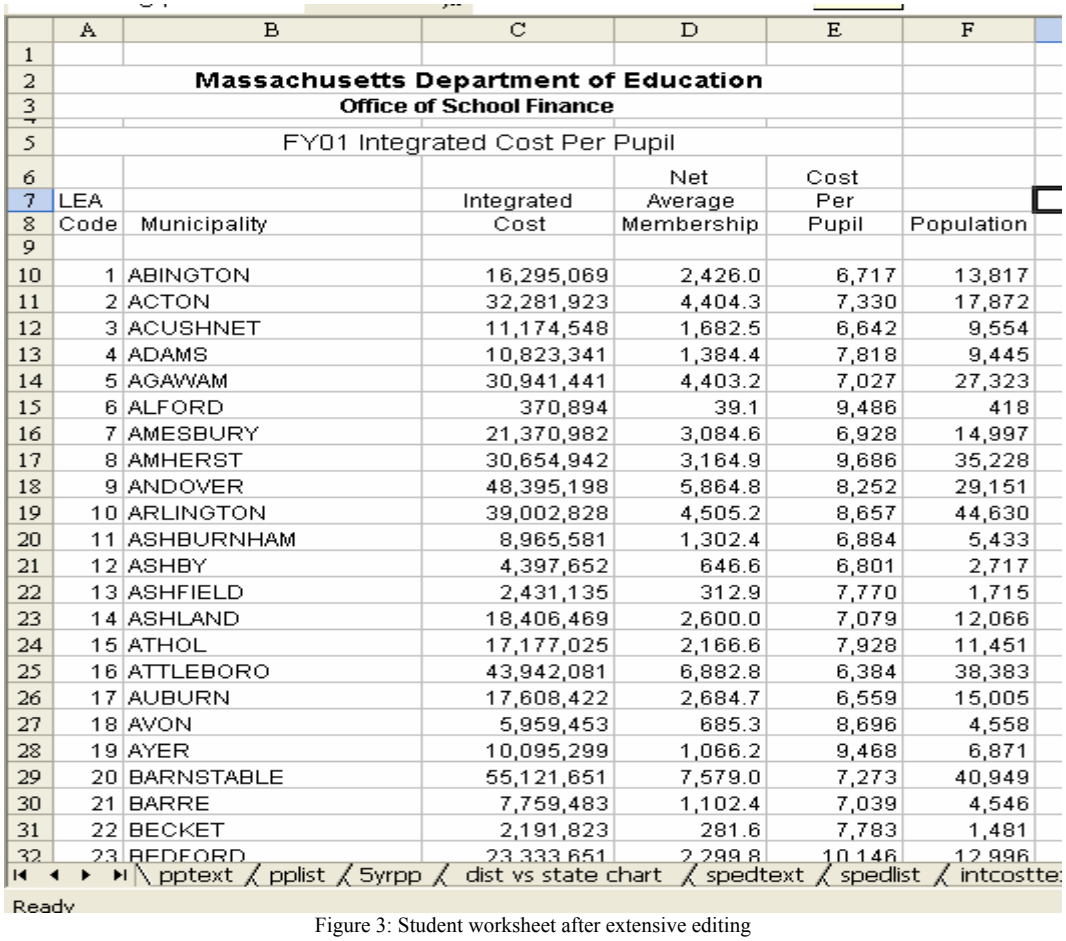
The data in the right hand column above (the copied data from MISER website in column G) are in correct collating sequence while the Department of Education data in the left hand column (column B) are not. This can be addressed by resorting individual portions of the worksheet, being careful to resort the original data columns (A through E) separately from the columns copied into the worksheet (G and H). Any errors that remain would be identified by the remaining "Out of sequence" messages in the logical test column. Facilitated by use of the Data Filter command, the student must deal with these remaining cases deal manually. These errors consist of footnote symbols, incomplete data (e.g., 'Gt. Barrington' instead of 'Great Barrington') and similar minor errors.

Now the student has the educational spending data from all 351 cities and towns on the same line as the population of those cities and towns. The student may now eliminate the town information in column G as redundant, as well as the spacer column in column F and the check column demonstrated above in column J. The result is shown Figure 3. Only now can the student begin the analysis assigned by the professor. The population data from MISER is now located in column F of the student's worksheet.

In no way do we wish to convey criticism of sites such as those used above. These sites are not constructed for the purpose of making available data for subsequent analysis. Rather they exist to convey information to their constituents. Indeed, in many cases they are not required to make the data available at all, and we should be grateful that our students benefit from their willingness to share data.

Often the site's workbook originated as an internal document, and thus there is no motivation to avoid internal jargon and assumptions. For example, as noted above, the student would likely use the 'pplist' worksheet, which has title 'Per Pupil Expenditure FY01', although in fact the needed information is actually located several worksheets away in a worksheet labeled 'intcostlist' and titled 'FY01 Integrated Cost Per Pupil'. These terms are likely to be meaningful to the Department of Education employees who created the workbook, and to the educational administrators across the state that are likely the most common users of this worksheet. However, the student is not likely to know that 'Integrated' means 'per pupil by city or town'. In fact, the student's knowledge of working with multiple worksheets is likely to be shaky at best. Moreover, the web site's owner's often have a strong desire to present the data in a highly readable format – hence the blank lines, summarizations and other factors which bedeviled us above.

Thus, even a site that offers data in Excel format often cannot be used directly by the student. We submit that it is questionable



	A	B	C	D	E	F
1						
2		Massachusetts Department of Education				
3		Office of School Finance				
5		FY01 Integrated Cost Per Pupil				
6				Net	Cost	
7	LEA		Integrated	Average	Per	
8	Code	Municipality	Cost	Membership	Pupil	Population
9						
10	1	ABINGTON	16,295,069	2,426.0	6,717	13,817
11	2	ACTON	32,281,923	4,404.3	7,330	17,872
12	3	ACUSHNET	11,174,548	1,682.5	6,642	9,554
13	4	ADAMS	10,823,341	1,384.4	7,818	9,445
14	5	AGAWAM	30,941,441	4,403.2	7,027	27,323
15	6	ALFORD	370,894	39.1	9,486	418
16	7	AMESBURY	21,370,982	3,084.6	6,928	14,997
17	8	AMHERST	30,654,942	3,164.9	9,686	35,228
18	9	ANDOVER	48,395,198	5,864.8	8,252	29,151
19	10	ARLINGTON	39,002,828	4,505.2	8,657	44,630
20	11	ASHBURNHAM	8,965,581	1,302.4	6,884	5,433
21	12	ASHBY	4,397,652	646.6	6,801	2,717
22	13	ASHFIELD	2,431,135	312.9	7,770	1,715
23	14	ASHLAND	18,406,469	2,600.0	7,079	12,066
24	15	ATHOL	17,177,025	2,166.6	7,928	11,451
25	16	ATTLEBORO	43,942,081	6,882.8	6,384	38,383
26	17	AUBURN	17,608,422	2,684.7	6,559	15,005
27	18	AVON	5,959,453	685.3	8,696	4,558
28	19	AYER	10,095,299	1,066.2	9,468	6,871
29	20	BARNSTABLE	55,121,651	7,579.0	7,273	40,949
30	21	BARRE	7,759,483	1,102.4	7,039	4,546
31	22	BECKET	2,191,823	281.6	7,783	1,481
32	23	BEDFORD	23,333,651	2,299.8	10,146	12,996

pptext / pplist / 5yrpp / dist vs state chart / spedtext / spedlist / intcostte

Ready

Figure 3: Student worksheet after extensive editing

that a casual student user of Excel would possess these skills.

3. SCENARIO 2: DATA AVAILABLE IN TEXT FORMAT

Here, we examine those sites that purport to offer data in a tabular format, but do not do so in a standard file format. This data must be transformed into a state in which a standard analysis tool like Excel can use it. Here we will divide our discussion into two parts:

Scenario 2a: Text Data Available in Tabular Format

Here we address those sites that provide data in a text format (*.txt). Data may sometimes be imported directly to the students worksheet. More often, a multi-step process is required:

1. Capture the data by selecting and copy-pasting the data to an intermediate holding area such an Microsoft Office's Notepad™;
2. Advisedly, saving the data in Notepad's native format, *.txt;
3. Take advantage of the analysis tool's text import functionality, such as Microsoft Excel's Text Import Wizard;
4. Edit the imported data as necessary; and
5. Save the data in Excel's native format, *.xls.

The following example will demonstrate what needs to be done in this scenario. Suppose a student taking a course on forecasting must analyze time series data. The Bureau of Labor Statistics (BLS) site is a good source of time series data. The home page for the BLS is:

http://stats.bls.gov/.

www.bls.gov Search | A-Z Index

BLS Home | Programs & Surveys | Get Detailed Statistics | Glossary | What's New | Find It! In DOL

Data extracted on: June 24, 2003 (03:02 PM)

Average Price Data

Series Catalog:

Series ID : APU0100702111

Area : Northeast urban
Item : Bread, white, pan, per lb. (453.6 gm)

Data:

Series ID	Year	Period	Value
APU0100702111	2001	M01	1.039
APU0100702111	2001	M02	1.053

Figure 4: First two lines of price of bread time series

At this site, say that the student selects "get detailed statistics" and then for price data he or she might select CPI – Average Price Data. Say that the student is interested in the time series for the price of a pound of white bread in the Northeast from 2001-2003. He or she would see the data shown in Figure 4.

To get to the point of analyzing these data the student would have to follow the sequence of steps noted above. A portion of the resulting data file is shown in Figure 5.

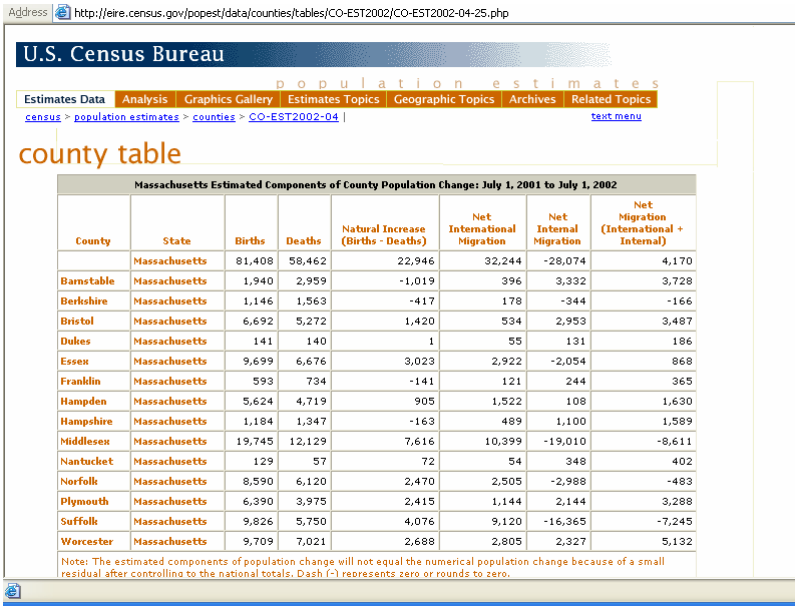
Series	Year	Period	Value
APU0100702111	2001	M01	1.039
APU0100702111	2001	M02	1.053
APU0100702111	2001	M03	1.12

Figure 5: Time series data for price of bread after importing into Excel

Even after completing the work, the data would likely need some editing. For example, the value column, representing the cost of a loaf of bread, would likely be reformatted as currency (although of course this is not necessary for the analysis itself). Similarly, the month code (e.g., 'M01') would likely be subjected to a table lookup to reflect the value 'January' or 'Jan', although, as it happens, the month data is continuous, and thus an autofill would suffice. . These steps may be superfluous for analyzing the data, but may be necessary for data presentation.

Scenario 2b: Data Available In Text Format But With A Layout That Inhibits Downloading

Here we address those sites that have data available in, at best, only roughly a tabular format, and do not purport to offer it for download. All of the issues discussed in Scenario 2a, above, apply here. Further, the numeric information discussed here is posted to the Web solely for the user to read, with no obvious intention of facilitating or even allowing the user to download the data for subsequent analysis. In this case,



Address http://eire.census.gov/popest/data/counties/tables/CO-EST2002/CO-EST2002-04-25.php

U.S. Census Bureau

population estimates

Estimates Data | Analysis | Graphics Gallery | Estimates Topics | Geographic Topics | Archives | Related Topics

census > population estimates > counties > CO-EST2002-04 | text menu

county table

Massachusetts Estimated Components of County Population Change: July 1, 2001 to July 1, 2002

County	State	Births	Deaths	Natural Increase (Births - Deaths)	Net International Migration	Net Internal Migration	Net Migration (International + Internal)
	Massachusetts	81,408	58,462	22,946	32,244	-28,074	4,170
Barnstable	Massachusetts	1,940	2,959	-1,019	396	3,332	3,728
Berkshire	Massachusetts	1,146	1,563	-417	178	-344	-166
Bristol	Massachusetts	6,692	5,272	1,420	534	2,953	3,487
Dukes	Massachusetts	141	140	1	55	131	186
Essex	Massachusetts	9,699	6,676	3,023	2,922	-2,054	868
Franklin	Massachusetts	593	734	-141	121	244	365
Hampden	Massachusetts	5,624	4,719	905	1,522	108	1,630
Hampshire	Massachusetts	1,184	1,347	-163	489	1,100	1,589
Middlesex	Massachusetts	19,745	12,129	7,616	10,399	-19,010	-8,611
Nantucket	Massachusetts	129	57	72	54	348	402
Norfolk	Massachusetts	8,590	6,120	2,470	2,505	-2,988	-483
Plymouth	Massachusetts	6,390	3,975	2,415	1,144	2,144	3,288
Suffolk	Massachusetts	9,826	5,750	4,076	9,120	-16,365	-7,245
Worcester	Massachusetts	9,709	7,021	2,688	2,805	2,327	5,132

Note: The estimated components of population change will not equal the numerical population change because of a small residual after controlling to the national totals. Dash (-) represents zero or rounds to zero.

Figure 6: US census data for Massachusetts by county

the only realistic way to capture the data is to select carefully the portion of the data to be used on the screen to be used, copy it and then paste it in the worksheet. In most cases, the data can be pasted directly into the analysis software. In some cases, as in Scenario 2a above, it must be pasted to a text editor such as Notepad for formatting before it can be effectively imported into the spreadsheet software.

By way of example, suppose a professor has assigned a student to prepare a graphical representation of population flows in the Commonwealth of Massachusetts by county. After searching the World Wide Web, the students finds the following site:

http://eire.census.gov/popest/data/counties/tables/CO-EST2002/CO-EST2002-04-25.php

which displays estimated 2002 data for birth, deaths and migration by county for the Commonwealth of Massachusetts (data for other states are available separately). Here, these data are simply displayed in tabular format, but no effort is made to make the data available in a standard file format. An example of these data is shown in Figure 6.

To download these data, the student must first highlight only the data portion of the table, avoiding header information (column headers, however, are acceptable). Note that longer column headers are "wrapped" over several lines, so the first line should be edited to incorporate the full column name. This action presumes that the student has the ability to:

1. go into the subsequent lines;
2. initiate an edit (F2);
3. cut the data in the line;
4. move to the first line; initiate an edit in that cell; and
5. paste the data on the clipboard to the end of the data in the first line.

The student would then have to repeat this for each column header that appears on subsequent lines. (Alternatively, the CONCATENATE function could be used, but we submit that the concept of concatenation, much less the use of the function, is beyond the capability of the casual student user.) Once finished, remaining blank lines between the column header and the data can be eliminated.

These data also present a very common problem: the handling of a calculated field. Here, there are two: "Natural Increase (Births – Deaths)" and "Net Migration (International + Internal)". We will use the later for discussion purposes.

"Net Migration (International + Internal)" is a calculated field because it is based on data held in the "Net International Migration" and "Net Internal Migration" columns. This column as downloaded, however, does not represent this but rather simply shows the sum as an entered value. A casual student user might not pick up on this distinction as s/he analyzes these data. As a result, the student would download this column and move it to Excel, as has been discussed above, but the column in the resulting worksheet would not capture any subsequent changes in the two columns to its left, "Net International Migration" and "Net Internal Migration". It is not difficult to predict that, in a short time, the "Net Migration" column would be in error because its underlying values in the two columns to its left would change over time. It is also not difficult to predict that the student would be unaware of this. It would be best if the two calculated fields were simply eliminated and recreated by the student as he or she requires the data, but we question if many students have that level of sophistication.

Further note that the first data line is a sum of the detail lines below, but is not labeled as such. We submit that many students would not catch this detail and thus their subsequent analysis would be skewed.

4. SCENARIO 3: DATA NOT IN TEXT FORMAT OR TABULAR LAYOUT

Here we address a more complex, but all too common, occurrence: websites that have data available, but in which the presentation of data is intended only for reading and in which the data are not in tabular form. Here the student must download the data as text (with the problems already discussed in Scenario 2). The student must then edit the data extensively in order to bring the data into a tabular format.

When the data is presented in a columnar format, the solution is straightforward. The student can generate an iterative macro that will move the data from a vertical orientation to a horizontal one on a cell-by-cell basis. The question then becomes: does a casual student user have the skill to write a procedural program (even without its iterative nature) and, if so, does s/he possess the ability to generate the procedural and iterative macro in VBA? Alternatively, in a more tabular format, the student could copy the data and then use the Transpose option of the Paste Special command, but again the question must be raised: is it reasonable to expect a casual student user to know about, much less utilize, this functionality?

This situation will be demonstrated by using an example. Suppose a professor has assigned a paper on the "Deaths Caused by Alcohol". A good deal of browsing has lead the student to the following site Department of Justice site

http://www.ojp.usdoj.gov/bjs/pubalp2.htm

where s/he selects a report on "Alcohol and Crime". On a subsequent screen the student is given the choice of downloading the file in *.pdf or *.txt format; he or she chooses *.txt.

Upon opening the file, the student finds that it is a textual report. However valuable that may be, the student needs statistical information for the assignment. Scanning down the report, the student finds summary statistical data, but it is integrated into the text. The question before us is: how will the student extract the needed data, and move it to Excel for subsequent analysis?

Let us begin with the occurrence of alcohol-related deaths (per 100,000 population) for all groups organized by five-year categories, the first table in the report. The data appear in Figure 7.

The data of interest are in the left most numeric column, and is a summary statistic ignoring the categorical distinctions of the columns to the right. To work on this data in a spreadsheet, one must copy the data to a text editor such as Notepad, and edit out extraneous data, if any. Next, the data must be saved as its own data file. (Any extraneous elements not eliminated here can be eliminated after moving the data to Excel.)

Next, the data must be opened in Excel, which will cause the Text Import Wizard to be invoked. This is shown in Figure 8.

```
            Rates of death caused by alcohol
            per 100,000 persons, adjusted for age

                                    Whites                Blacks
Year        All groups    Males    Females    Males     Females
1980          8.4         10.8      3.5        32.4      10.6
1985          7.0          9.2      2.8        27.7       8.0
1990          7.2          9.9      2.8        26.6       7.7
1992          6.8          9.9      2.6        22.3       6.3
1993          6.7          9.7      2.7        21.3       5.5
1994          6.8          9.9      2.7        20.4       5.6

Percent change
1980-94     -19.0%       -8.3%    -22.9%     -37.0%     -47.2%

Source:  National Center for Health Statistics, Monthly Vital Statistics
Reports.
-------------------------------------------------------------------------
```

Figure 7: Straightforward data download via a text editor

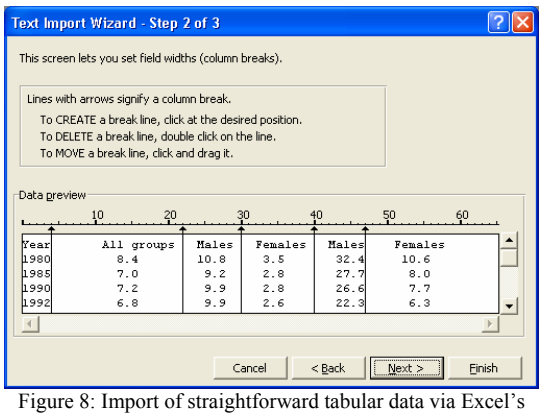


Figure 8: Import of straightforward tabular data via Excel's Text Import Wizard

If not done previously, the student may then "clean up" the data by eliminating unneeded columns, and proceed with the analysis. Thus, the download of these data and their transformation to Excel format is quite straightforward.

More challenging, however, are data tables that have been edited for display purposes in a document. This editing is entirely legitimate in itself, but if the data are not provided in tabular form, considerable "massaging" of the data is necessary before students can use it in a standard data analysis tool.

This can be seen clearly with Table 16 on this site, which presents traffic fatalities and data relating to alcohol involvement for the fifty states and the District of Columbia. However, it is presented in two columns for editing and space-saving reasons. These data are displayed in Figure 9.

The challenge here is a data file with four fields that is doubled horizontally and must be downloaded as a data file of eight fields. The first task is to save the left hand data as a text file, and import it into Excel. However, the title material in the first several lines causes the Text Import Wizard make incorrect assumptions as to data delimitation. (This could be solved by eliminating the titles from the imported data.) Here, the student will have to switch from fixed width to specific character delimitation. By experimenting, the student finds that a space delimiter will work best, with a key exception. Spaces in state names (e.g., 'New—Hampshire') will be treated as delimiters. For the data as a whole, there are eleven cases like this, from 'District—of—Columbia' to 'West—Virginia'. Here, they must be fixed manually, as must other minor errors. Alternatively, as noted in Scenario 2b, above, the CONCATENATE function could be used.

Then, the student must repeat the process for the twenty-six states to the right, after which these data must be cut and pasted to the bottom of the twenty-five states to the left of the table, so that these data are in the rows below the first group of twenty-five states.

A common problem now appears. The integer field 'Number of fatalities' is fine, but the remaining two columns are percentages. As they stand now, they do no harm, but the student's subsequent data analysis may well require their use a multipliers. For example, the student may need to report the *number* of alcohol-related fatalities (that is, 'Number

Figure 16.

	Number of fatalities	Percent of fatalities Involved alcohol	Drivers had a BAC 0.10 or higher		Number of fatalities	Percent of fatalities Involved alcohol	Drivers a BAC 0. or higher
U.S. total	41907	40.9%	32.0%	Missouri	1149	49.4	38.7
Alabama	1143	42.6	34.1	Montana	200	37.1	33.0
Alaska	80	51.1	44.4	Nebraska	293	33.6	26.0
Arizona	993	43.9	34.8	Nevada	348	50.1	37.3
Arkansas	615	34.7	26.9	New Hampshire	134	34.7	26.4
California	3989	40.2	30.1	New Jersey	818	34.2	25.4
Colorado	617	39.6	33.0	New Mexico	481	50.1	42.0
Connecticut	310	49.2	38.2	New York	1564	33.4	24.0
Delaware	116	41.0	28.5	North Carolina	1493	35.1	27.8
District of	62	49.2	36.0	North Dakota	85	53.4	44.9
Florida	2753	36.9	29.1	Ohio	1395	33.0	25.6
Georgia	1574	36.0	27.6	Oklahoma	772	36.3	28.2
Hawaii	148	44.4	31.5	Oregon	524	42.2	32.6
Idaho	258	33.8	26.0	Pennsylvania	1469	39.1	32.3
Illinois	1477	45.0	36.1	Rhode Island	69	48.4	36.2
Indiana	984	34.1	27.0	South Carolina	930	42.4	33.6
Iowa	465	42.5	33.0	South Dakota	175	39.9	31.1
Kansas	491	40.9	30.0	Tennessee	1239	40.2	32.8
Kentucky	841	35.2	28.1	Texas	3741	53.2	42.3
Louisiana	781	51.4	39.6	Utah	321	23.7	18.7
Maine	169	37.5	29.0	Vermont	88	43.9	36.9
Maryland	608	33.0	24.0	Virginia	875	38.6	30.2
Massachusetts	417	44.4	32.0	Washington	712	50.0	40.1
Michigan	1505	40.7	31.6	West Virginia	345	38.0	31.6
Minnesota	576	37.9	30.0	Wisconsin	761	42.4	34.2
Mississippi	811	41.6	33.1	Wyoming	143	40.6	26.2

Source: Alcohol Traffic Safety Facts 1996, National Highway Traffic Safety Administration.

Figure 9: Download difficulties in text files and in non-tabular format

of Fatalities' **x** 'Involved Alcohol'). Using the first data line (Alabama) as an example, a student would find that Alabama had (41,907 **x** 40.9 =) 1,713,996 traffic fatalities involving alcohol, or 38.5% of the state's population! Such an error would escape many students. As a result, the student must be store the percentage values as proportions and then format them in percentage format or percentage style. To accomplish this, the student must:

1. Create a new column beside each percentage column;
2. Divide each value by 100 to express it as a proportion;
3. Copy-Paste Special each cell as a value (that is, to divorce the value from its source);
4. To eliminate the original percentage data; and
5. To reformat the proportion data as a percentage.

Thus, before these data can be analyzed in Excel, there must first occur all of the steps of Scenario 2 above. In addition, the student must perform on the data a significant amount of data repositioning, computation, editing, alteration of the data ranges from computations to stored values, and formatting. The key question, as has been addressed above, is: can we expect a casual student user of Excel to be capable of doing this?

5. IMPLICATIONS FOR FACULTY AND CONCLUSIONS

Faculty, particularly those outside the IS discipline, often assign students to "get data from the web". Our intention has been to demonstrate that accomplishing this seemingly straightforward task can be quite challenging, particularly from the perspective of a casual student user of a data analysis tool such as Excel. Even those sites that offer data downloads in Excel's native format pose serious difficulties. These difficulties increase significantly when the data available must pass through a text format before being imported into Excel.

We submit that many professors do not understand these difficulties and that their assignment to "get data from the web" represents to students a more challenging task than the professor had in mind. Moreover, the time that students spend in downloading, importing and editing data prior to analysis inevitably reduces the time they spend on the assignment itself, thus reducing the educational value of the assignment. In fact, the present authors have observed students manually retyping data found on the web into Excel – surely

this is a waste of time for the student, with no educational value whatsoever.

We advocate a two-part program to address these difficulties:

1. We recommend that professors make it a habit to test how difficult it is to "get data from the web" prior to making the assignment. We intuit that many professors have very limited experience in doing this, or if they have, they have limited their use to viewing the data, and not downloading the data for subsequent analysis. Performing the necessary data downloads prior to giving the assignment will allow the professor to assess how realistic it is for students to download the data themselves, given their skill set.
2. While the experience of downloading data from the web has educational value, we recommend that the professor limit it to a few assignments. After that, we recommend that the professor or a teaching assistant go through the download process and make the data available to the student in a correct format for the data analysis tool to be used.

Our findings have implications for colleges and universities as well. First, faculty must educate administrators of the difficulties students face when downloading data from the web. Second, administrators must make available to students support in using the web to find and download data. They must recognize that downloading data is beyond the capabilities we can reasonably expect from casual student users. This support could be done straightforwardly from existing help desk organizations, provided that the works that staff that function are trained in the issues above.

Finally, our findings have implications for the IS community. Data warehouse issues have, thus far, been taught conceptually, but we predict that technical instruction will soon be available. When that occurs, IS faculty will face much the same difficulties as described above for faculty outside of IS. By definition, a data warehouse is a "repository of information collected from multiple sources" (Han, 2001). Moreover, the data mining's definition includes query, analyzing and interpreting the data. While a data warehouse is defined to include a common and unified data schema, in practice this often means simply that data are held in a common storage medium, but in their native file formats.

To what degree the data warehouse software can itself resolve these issues, either at the time the data is stored or when it is retrieved for a specific analysis, is unknown. Thus, IS instructors could well face the same difficulties described earlier for today's non-IS instructors: the need to capture data in multiple formats for analysis under a common analysis tools.

6. REFERENCES

Gorgone, John T, *et al*., eds., (2002). IS 2002 Model Curriculum and Guidelines for Undergraduate Programs in Information Systems, Association for Information Systems.

Han, Jiawei, and Micheline Kamber, (2001). Data Mining: Concepts and Techniques, Morgan Kaufmann, p. 12