

# Data Links

TriUniversity Data Resources

tdr.tug-libraries.on.ca

Vol. 4 / Issue 2 Winter 2001

Editor: Shabiran Rahman • [srahman@library.uwaterloo.ca](mailto:srahman@library.uwaterloo.ca)

## **TDR HOME PAGE HAS A NEW LOOK**

Some of you may have noticed the recent 'cosmetic' changes to our web pages. In late December the layout and organization of our web pages changed significantly. We moved to using rollovers and changed the colour scheme and organization of information. A great deal of work went into deciding how best to present our data and how to use some more up to date web tools. The general consensus has been very positive. The process is, however, ongoing and many new tools and help files will be added over the next few months.

Possibly more significant is what happened behind the scenes. From the user perspective our retrieval system has moved to a newer and much faster Unix system. This has allowed for significant improvements in retrieval times as well as allowing us to mount larger data sets with millions of observations (such as the LFS and the World Trade DataBase). The new system has also given us expanded disk options and our collection is now close to 100 GBs with room for another 100 GBs.

## **CONTENTS**

- TDR Home Page Has a New Look
- How to Prepare Web Retrieval Data for Statistical Analysis
- Featured Data Set: World Trade Database
- What is IASSIST?
- The Southwestern Research Data Centre
- Pat Newcombe-Welch Appointed Statistics Canada Analyst
- New Acquisitions
- Sites of Interest

## **HOW TO PREPARE WEB RETRIEVAL DATA FOR STATISTICAL ANALYSIS:**

### *A step by step account*

In the last issue of Data Links, I presented a summary of resources for helping with the analysis of statistical data at the three universities. In this article, we will look at an overview of how to prepare the data retrieved by the web retrieval system for analysis by a statistical package.

To illustrate, let us look at an example using the 1997 Survey of Consumer Finances for Individuals. This survey contains information on income and labour-related characteristics of individuals. Suppose I wish to analyze several characteristics of Ontario income earners in 1997. In particular, I want to examine the relationship of several factors on total individual income for individuals who earn more than 50% of their household's income. In this example, I wish to use SPSS to perform the analysis.

First, I need to select the proper choices on the data retrieval page for the SCF-IND 1997 data set:



I've entered a Unix ID and selected the option to Keep Information for seven days. I do this if, for any reason, I need to re-retrieve the data while I am preparing it for analysis.

In the prov subsetting box, I've chosen Ontario in order to restrict data retrieved to only Ontario individuals. (I could have done this later on by a different method, but doing this now on the web retrieval form saves a lot of work.)

I've chosen a set of variables I think I will need for my set of analyses in SPSS:



To be safe, I've chosen a larger group than I will probably use so I won't have to return here and do another retrieval. I've chosen SPSS for Windows as the Output type, and I've specified a listing file to be created. Let me explain these last two choices in some more detail.

Would I still have been able to analyze this data in SPSS if I had kept the default Output Data Type (ASCII - comma delimited)? The answer is yes. If I had saved my data as an ASCII file, I would still be able to analyze it in SPSS using the import data feature, which would convert my data from ASCII format to the internal format used by SPSS. The advantages of specifying SPSS as the output type is that the labels are (somewhat) kept intact, and the variable names are automatically preserved. Also, by saving the output as an SPSS data set, the data can be immediately read by SPSS. However, the disadvantages to saving as an SPSS file instead of an ASCII file are that there are some glitches in the automatic conversion process—not all information about the variables and labels is automatically preserved in the SPSS file. By saving the file as ASCII and then importing it into SPSS, I will be required to perform more manual work, but I will also be more assured of the integrity of the data and variable information.

The creation of a listing file is for verification purposes only. The listing file is not necessary for creating the output file—the output file is created independently of the listing file.

In my example, I have chosen the following variables: Age grouping (AGEGRP), total income code (INCCODE), individual income (TOTINC), % of household income (PROPINC), sex (SEX), household type (HHTYPE), immigration status (IMMIGSTA), marital status (MARSTAT), class of worker (CLSWRKS), industry (INDUSTRY), and education level (EDUCREC). Please note that not all of these variables are visible on the screen used for this example. When you work with this, you will be able to scroll up and down to make the required selections.

The next step is to open the saved file in SPSS and perform some massaging of the data to facilitate data analysis. The amount of data massaging necessary depends on the kinds of analyses to be performed. The main types of data massaging will consist of data transformation, recoding, and data set subsetting.

Data transformation is the process of creating new variables or modifying existing variables based on data set values. In my example, I may wish to create a new variable to be the natural logarithm of TOTINC. Recoding is the process of transforming the values of a class variable to a different set of values. For example, I may wish to recode the

values of SEX so that "1"="M" and "2"="F". Data set subsetting is the process of selecting a subset of data for analysis based on certain criteria. For example, I may wish to generate statistics for individuals whose proportion of total household income is greater than 50%.



Other types of massaging may be performed to tailor the data to a certain type of analysis. Generally, you will find it easier to alter the data after you have determined the type of analyses you wish to run. It is at this point that consulting with your local statistical consultants and data analysis experts can help you achieve the desired results.

Jack Cooper, jack@ist.uwaterloo.ca

### **FEATURED DATA SET: WORLD TRADE DATABASE**

Since our last newsletter several dozen new data sets have been requested and added to our collection. One of the most significant is the World Trade DataBase. It contains detailed data that was previously unavailable in our hard copy collections, and as such is in high demand.

This is an enormous data set that provides information on imports and exports between Canada and all the countries of the world. There are over 40 million records in total. The data is reported monthly in terms of both dollar values and quantities. The information is broken down to the 10 digit SITC (rev3). This is extremely detailed. As an example, one product is specifically defined as "Degreased shorn wool, not carded, combed or carbonised." Because of this detail, the data set is very large and is divided into several sections. The enormity of the data can sometimes make it a little difficult to use the first time. If data at the 2, 4, 6, or 8 digit SITC is needed, one needs to aggregate using the mean's feature. We recommend that you ask your local support staff to walk you through the intricacies of the data set the first time. Several other sources of trade statistics are also available and compliment this collection. These include such things as OECD - International Trade by Commodities, IMF Direction of Trade, and Industry Canada's excellent strategies site on trade.

Bo Wandschneider, bo@uoguelph.ca

### **CONGRATULATIONS**

Congratulations to Susan Moskal of UW Electronic Data Service who received the Ontario College and University Library Association award at the OLA 100th Anniversary Super Conference in Toronto

[www.accessola.com/superconference2001/](http://www.accessola.com/superconference2001/)

### **CAPDU MEETING**

The next meeting of the Canadian Association of Public Data Users <http://www.ssc.uwo.ca/assoc/capdu/> will be held at Université de Montréal, April 26, 2001 – April 28, 2001

Membership is \$25 If you are interested in becoming a member of this association , please fill out the form at [www.lib.uwaterloo.ca/TUG/gp/publicdata.html](http://www.lib.uwaterloo.ca/TUG/gp/publicdata.html) and send your check to

Shabiran Rahman, Treasurer CAPDU,  
University of Waterloo,  
200 University Avenue W,  
Waterloo, Ontario, N2L 3G1

---

### **WHAT IS IASSIST?**

IASSIST is an acronym for the International Association for Social Science Information Service and Technology. It brings together professionals from around the world to help in the promotion of social science research. It is an organization dedicated to the issues and concerns of data librarians, data archivists, data producers, and data users. Several members of the TDR have attended and presented at recent meetings held at Yale, Toronto, and Northwestern. Sessions included such topics as "Working with Census Data in Arcview, SAS, SPSS, and Stata," "Preparing Data for the

User Community," "Hyper Linking the World of Social Science: Integrating Text and Data in a Global Hypertext Space," "Research Data Centres and Confidential Data" and "Promoting Use of Numeric Data Sets in Learning and Teaching Through Enhanced Local Support." The list is extensive.

Some of the main goals of IASSIST are:

- Professional development of staff in social data information centres
- advancement and development of social data information centres
- assessment of and planning for the impact of new technology
- promotion of the archiving of social data and the advancement of data standards
- promotion of global linkages between social data centres
- development of linkages between social data centres and users and producers of data including the academic, public, and private sectors; evaluation of the role and contribution of IASSIST
- the recruitment of new members.

For more information see:

[datalib.library.ualberta.ca:80/iassist/](http://datalib.library.ualberta.ca:80/iassist/)



---

**Pat Newcombe-Welch**  
**Appointed Statistics Canada Analyst**

Pat Newcombe-Welch, a survey methodologist from the Social Survey Methods Division of Statistics Canada, has been appointed as the Statistics Canada analyst at the Southwestern Ontario Data Research Centre, scheduled to open at UW in late spring.

Pat, who holds a Ph.D. in statistics (UW, 1994) has worked as a statistical consultant in the Department of Statistics and Actuarial Science for the past three years, while on family related leave from her position in Ottawa. A native of Southwestern Ontario, Pat was born and raised in St. Thomas, and completed her B.Sc. and M.Sc. degrees at the University of Guelph.

---

## **THE SOUTHWESTERN RESEARCH DATA CENTRE**

The Southwestern Ontario Research Data Centre (RDC), scheduled to open in late spring at UW, is one of the results of the Canadian Initiative on Social Statistics, a joint project of the Social Sciences and Humanities Research Council of Canada (SSHRC) and Statistics Canada.

A National Task Force made up of leading Canadian researchers and statisticians was formed in 1998 in order to address the need for a "national capacity to fully analyse" the "rich and unique set of data collection instruments and data sets that Statistics Canada has developed in recent years."

A discussion of the Canadian Initiative on Social Statistics can be found at [www.sshrc.ca/english/policydocs/discussion/statscan.html](http://www.sshrc.ca/english/policydocs/discussion/statscan.html) and the final report of the Task Force can be downloaded free of charge by following the instructions at the end of that discussion. Information about the RDC to be housed at UW can be obtained by going to the UW Survey Research Centre home page at [www.stats.uwaterloo.ca/Stats\\_Dept/SRN/](http://www.stats.uwaterloo.ca/Stats_Dept/SRN/) and clicking on the Southwestern Ontario Research Data Centre.

Pat Newcombe-Welch  
[panewcom@setosa.uwaterloo.ca](mailto:panewcom@setosa.uwaterloo.ca)

---

## **A FEW OF OUR NEW ACQUISITIONS**

The following datasets have been added to the collection as of January, 2001

Absence from Work Survey 1999  
Canadians Out-of-Employment Survey 1995  
ICPSR General Social Surveys 1972-1998 (U.S.) - Cumulative File  
International Travel Survey 1990-1998  
National Population Health Survey 1998-99 Cycle 3 Households - General & Health files  
International Travel Survey 1990-1998  
National Population Health Survey 1998-99 Cycle 3 Households - General & Health files

---

### **SITES OF INTEREST FOR YOU TO EXPLORE:**

In this segment we will be presenting the URLs of one or more sites that may be of interest to data users. Below are our picks for this issue. Both these sites were recommended by Dr. John Wilson, University of Waterloo and would be of interest to students of election studies.

If you would like to recommend a site for inclusion in this segment, please send it to:  
Shabiran Rahman [srahman@library.uwaterloo.ca](mailto:srahman@library.uwaterloo.ca).

LISPOP (Laurier Institute for the Study of Public Opinion and Policy) [www.wlu.ca/lispop](http://www.wlu.ca/lispop)

University of Waterloo Centre for Election Studies <http://www.arts.uwaterloo.ca/PSCI/ces1.htm>

### **YOUR CONTACTS FOR DATA SERVICE AT OUR THREE LOCATIONS**

#### **University of Guelph**

Bo Wandschneider  
[bo@uoguelph.ca](mailto:bo@uoguelph.ca)  
519-824-4120 x6410

#### **Wilfrid Laurier University**

Helene LeBlanc  
[hleblanc@wlu.ca](mailto:hleblanc@wlu.ca)  
519-884-0710 x3743

#### **University of Waterloo**

Susan Moskal  
[srmoskal@library.uwaterloo.ca](mailto:srmoskal@library.uwaterloo.ca)  
519-888-4567 x2890

Shabiran Rahman  
[srahman@library.uwaterloo.ca](mailto:srahman@library.uwaterloo.ca)  
519-888-4567 x2882