Accessing data through Wharton Research Data Services' web interface

Wharton Research Data Services (WRDS) is a data management system that enables faculty and students to access some important research datasets through a web interface. Kellogg does not have access to all the datasets listed in WRDS; it can only access those datasets to which it subscribes: all the CRSP data, Standard & Poor's Compustat (North America, annual updates) and ExecuComp, and TAQ are the most important. Kellogg does not have access to S&P databases labeled as "Global" in WRDS, for example. If a Kellogg user submits a query to a resource to which the school does not subscribe, the system will return an error message ("nwu does not have permission to access…"). Access to these data is solely for academic research.

Full descriptions of Compustat and CRSP, in addition to other datasets available at Kellogg, are available in the Research Computing web site:

www.kellogg.nwu.edu/researchcomputing/

This site includes also online documentation to datasets and software, when available.

Keep in mind that all NU and Kellogg subscriptions are for academic (research / educational) use only. Commercial and for-profit use, such as in the context of an internship or job, is prohibited by our licensing agreements.

Logging-in to the WRDS web site:

For MBA students, there are two ways of accessing the WRDS server:

- With a class login and password provided by a faculty member (only Kellogg faculty may apply for class accounts), point your browser to:

  wrds.wharton.upenn.edu

  To login, click on "members login". Use the ID and password provided by your instructor.

- For authentication based on a computer's IP address, point your browser to:

  wrds.wharton.upenn.edu/connect

  This method works within the Jacobs Center, and in the computer lab (as well as library workstations) in Wieboldt Hall.

Once authenticated, WRDS will open the "home" window, which lists the available data collections (also available through the pull down menu at the top of the screen). You can return to this screen anytime, by clicking on the "Home" button at the top.

On the left side of the screen and in a toolbar across the top, there is a link to WRDS support area, which includes a section on Frequently Asked Questions (FAQ), manuals, downloads, and links to data vendor sites. Of these links, the link to the manuals is most likely the only one relevant for users of the web interface. Many of the FAQs deal with

issues relevant to users of WRDS' Unix host. WRDS' Forums can also be searched, but class accounts and IP logins are not allowed to post on them.

Pages available for each dataset

From the WRDS' home page, clicking on a data collection or selecting it from the pull-down menu will take you to the same, introductory, page on the dataset. This page provides a description of the dataset (or collection of datasets). Both CRSP and Compustat are collections of datasets. Thus, on the left, you will see links to each of the datasets within the collection.

Once you select a specific data collection, you will get a description. On the left side, you will see a list of the different datasets that make the data collection. There are three pages related to any given dataset:

- Query page
- Documentation page
- Data Manuals page

When you select a specific dataset (for example, the "Industrial Annual" file), the first screen WRDS will display is the query screen.

For the most part, you are going to use the Compustat Industrial file, and CRSP's Stock files (daily and monthly).

Query screen

This is the first page displayed. It is an HTML form that allows you to create a subset or extract from the dataset, by providing certain inputs.

In CRSP and Compustat it is frequent to extract information on certain companies. Individual companies are identified by a code that can be the CUSIP number (called CNUM in Compustat or CUSIP in CRSP), its stock exchange symbol (called SMBL in Compustat or TICKER in CRSP) or the datasets unique identifier (GVKEY in Compustat, PERMNO in CRSP, NPERMNO in the merged CRSP/Compustat database).

Alternatively, groups of companies can be selected according to their industry or the value of a specific variable (the latter option not available in the CRSP stock files). For example, in Compustat you could select companies with more than 3,000 employees.

In the query screen you must also select the variables you need, and the type of output file. The last section of this document offers to examples of data extraction

Documentation

The documentation screen provides a brief (one line) description of each of the items available in the query screen. For detailed information on each variable, check the data manuals.

Data manuals

The data manuals page provides detailed information about the specific dataset, when such documentation exists in electronic format. This is certainly the case for CRSP and Compustat, which offer their documentation in Acrobat PDF format. The same documentation is reproduced in the Research Computing web site.

For Compustat, the key document is chapter 5, "Data Definitions", which is an alphabetical list of the items included in the dataset. In many cases, the brief description in the documentation screen will be enough. In other cases, for example, you will need to refer to the data definition in the manuals. For example, if you need to decipher the codes (from 1 to 3) for the "S&P Index Primary Market" (variable name CPSPIN), you will look for this variable's definition in chapter 5 of the Compustat manual. If you cannot locate a specific variable in chapter 5, chapter 8 has cross-reference tables that list the variables according to their mnemonic and provide their "long" name in the data definitions chapter.

In CRSP, the data documentation is in a single PDF file per dataset. In this case, the key chapters in each file are the chapters on "Database Structure" and the "Data Definitions" chapter. Keep in mind that the CRSP manuals are particularly hard to follow as they have been written specifically for use with the original versions of the files, designed for Fortran and C access.

The CRSP manuals on the Research Computing page have been split into chapters, so you may click directly on the chapter of interest.

Some notes about the output options

The best options are a comma-delimited text file – a "CSV" file (comma-separated values) or a tab-delimited file, both formats easily read by any statistical package or spreadsheet.

The default is fixed-width text file. While any statistical package is able to read this type of file, it unnecessarily complicates the task at hand. The data starts at row 6 (with headers at row 4) and missing values are not acknowledged by a "place-holder". Thus, to read this type of output in a statistical package, the user would have to specify the column location for each variable, and each variable's format (in addition to making the software read starting at row 6).

In Stata, for example, to read a tab-delimited file, the command is

insheet using *filename*

SPSS users may make the appropriate selection in the File menu or write the syntax for a "get data" command.

Similarly, to read a CSV file, the Stata syntax is:

insheet using *filename*, comma

SPSS users may use the menu options or the "get data" command.

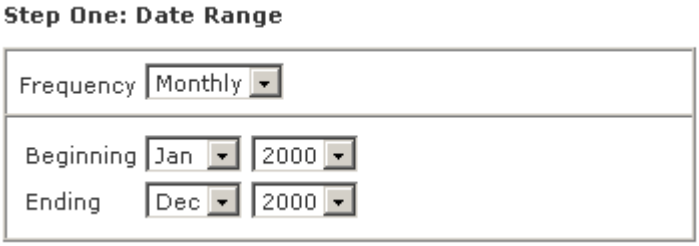
Examples of data extraction with the web interface

The first of the three examples that follows provides details on each step involved. Subsequent examples provide details only in those features in the database or web interface that differ with respect to the first example.

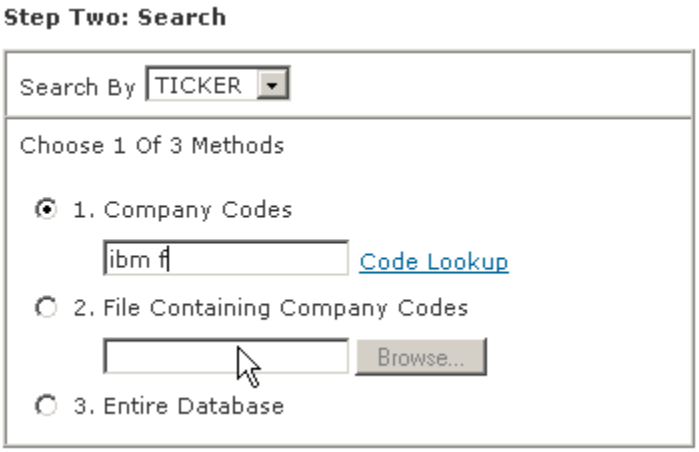
Example 1: Getting data on a IBM and Ford Motor Co. stocks (CRSP US Stock database)

This example outlines the steps you would follow to download monthly data (January through December 2000) for IBM and Ford, selecting the following variables: Price, holding period return and number of shares outstanding.

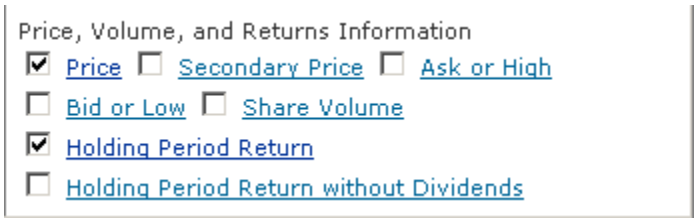
1. From the main page, select CRSP. This will bring up a general description of the CRSP collection of databases. On the left margin, select "Stocks", which will display the CRSP Stock Data Query Screen.
2. In "Step 1: Date Range", set the frequency to "monthly". Set the beginning month to January 2000 and make sure the Ending date is December 2000.



3. In "Step 2: Search" you have three options for searching the database. You may type a company code or provide a text file with a list of company codes; you may also select certain variables (Step 3) and download the information for all the securities available in the database. The "company code" can be the stock exchange ticker, its CUSIP number, the "permanent number" (PERMNO) assigned to the security by CRSP or the SIC code assigned to the company in the database (HSICCD). For this example, select option 1 and type the tickers for IBM (ibm) and Ford Motor Co. (F).



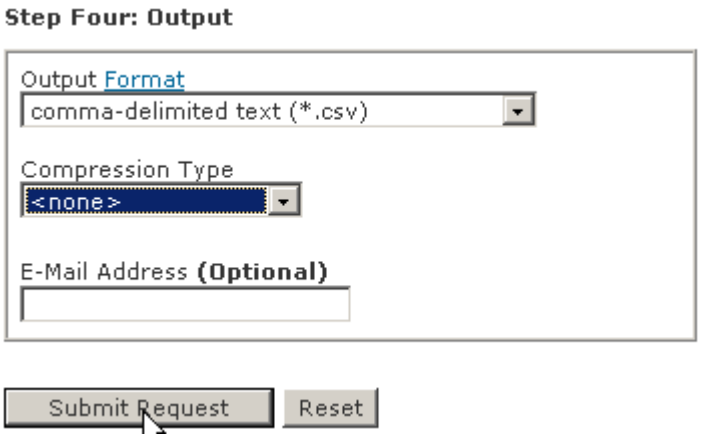
4. In "Step 3: Variables" you will notice that the variables are grouped into eight panels. You may click on any variable to get a description. For several of these variables (e.g., SIC codes, traits codes, delisting codes, etc) you will have to refer to CRSP's documentation to understand the meaning of the codes. For this example, we will select price and holding period return from the "Price, Volume, and Returns Information" panel.



To select a variable, click on the box to its left. This will place a checkmark in the box. To deselect it, click again on the box. To start the whole process over, click on the "Reset" button at the bottom of the screen.

Select "number of shares outstanding" in the "Share Information" panel:

5. Scroll down to "Step 4: Output". Select "comma-delimited text (*.csv)" from the drop down menu. Depending on the size of your request, you may also select to compress it. Any of the compression types listed in the compression drop down menu can be expanded with the WinZip utility. If the request is large and you need to logout, you may provide your e-mail address to be notified when the extract is complete.



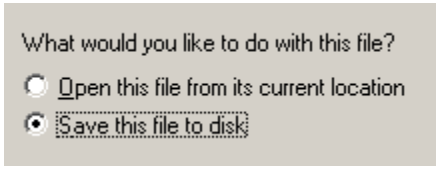
6. Before clicking on the "Submit Request" button, quickly verify your selections. Once you submit, a new browser window will open, which presents you a summary of the query being executed. This window will refresh every 10 seconds until the query is complete. Notice that after the query summary, the page includes a link to the page where the output will be placed. If the query is taking longer than you expected and you must logout, copy the URL generated by your query. In this example, the page generated by WRDS is shown below.

   Your request is being processed. When finished, the output will be found at
   http://wrds.wharton.upenn.edu/output/081417002.html

   Once the query is done, the URL will be replaced with a file name, as shown below:

   081417002.csv (1 KB)

   If you are using **Internet Explorer** and click on the file link, IE will provide two options:

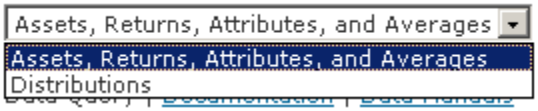
   If you choose to open it from WRDS, it will be displayed in an Excel viewer.

   If you are using **Netscape**, clicking on the file will show the data on the screen. You can save it to disk or, in the previous screen, right-click on the file's name and select "Save link as" from the pop-up menu to save the file to disk.

Example 2: Extracting data on Dimensional Fund Advisors funds (CRSP Survivor-Bias Free US Mutual Fund database)

This example selects DFA funds, extracting data on its asset composition (percentages invested in stocks, bonds and preferred stocks, preferred stocks, and cash equivalents) and monthly returns per share.

1. From the main page, select CRSP. On the left margin, select “Mutual Funds”. This will display the CRSP Mutual Funds Data Query Screen. If your browser was already pointing to a CRSP data query screen, click on “Mutual Funds” in the left margin.
2. The first step is to select the appropriate file. At the top of the CRSP Mutual Funds, make sure that “Assets, Returns, Attributes, and Averages” is selected.



3. In “Step One: Date Range”, select 1995 through 2000.
4. In “Step Two: Search” you can search funds by an identifier called “ICDI”. This is a code, generally numeric, that identifies the funds (albeit not uniquely). If the code starts with an “M”, it was assigned by CRSP and the data corresponds to a “dead” fund. To find the DFA funds, click on “Code Lookup”, and search for “dfa” on the form that will open. A list of matching funds will be shown.

ICDI	NAME	DATE RANGE
02092	DFA Invest Grp:US Large Cap Value Portfolio	1993-1997
02092	DFA Invest Grp:US Large Cap Value Port	1998-2000
03162	DFA Invest Grp:Pacific Rim Small Company Port	1993-2000

5. For this exercise, we will retrieve data for ICDI 19300. Type “19300” in the “Company Code” field. If you would like to retrieve all the funds displayed, you can select the table with the mouse and paste it into Excel. With the codes in a spreadsheet, you can then prepare a text file that has one ICDI code per line and can be read by the WRDS web interface.
6. In “Step Three: Variables”, the variables are grouped in five panels. Unlike the CRSP Stocks query screen, you cannot click on the variable name for a description. Instead, you must rely on the brief description provided under the “Documentation” link or the data manuals.

   Note that in the “Descriptive Information” panel the earliest and latest dates for which there is data available for a given fund are selected by default (you may deselect them). For this example, in the “Descriptive Information” panel select “Fund name”. In the “Asset Composition” panel, select the percentages invested in stocks, bonds and preferred stock, preferred stock, and cash equivalents. In the last panel, “Monthly Series”, select “total return per share”.
7. In “Step Four: Output”, select “comma-delimited text (*.csv)” and click on the “Submit Request” button if your query is ready. The resulting file will show one

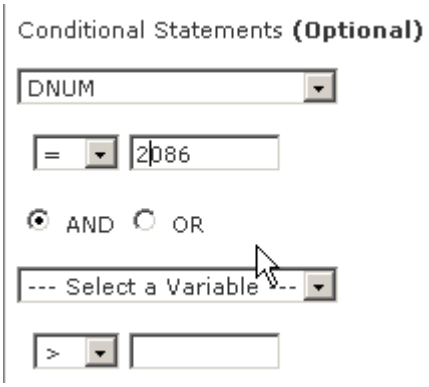
line of data per year (each of the monthly returns will be a column) until 1999. For 2000, the data for the first two quarters is shown.

Example 3: Extracting data for the soft drinks industry from Compustat (Compustat Industrial file)

When you click on the "Compustat" link in the WRDS site, the page you will see has a list on the left side. Please note that we do not subscribe to the quarterly updates. Also, note the "Tools" among those links, which allow the download of financial statements.

This example uses the "Conditional Statements" option available in the Data Query interface for Compustat. Using this option, the example downloads net sales (data12), income before extraordinary items (data18), and number of employees (data29), for the firms in the "Bottled and Canned Soft Drinks and Carbonated Waters" (SIC code 2086), for the period 1995-1999. The SIC code in Compustat is called "DNUM".

1. From the main page or the top of the screen, select "Compustat". On the left margin, among the "Annual Update Files", select "Industrial". Kellogg does not subscribe to the Compustat quarterly updates. (Quarterly data is available, but the files are updated once a year)
2. In "Step One: Date Range", make sure the frequency is set to annual and limit the range to 1995 through 1999.
3. In "Step Two: Search", select the third option, "Entire Database". In the drop-down menu that follows, make sure "Active and Inactive Companies" is selected. In the "Conditional Statements" panel, select "DNUM" from the first drop-down menu, the "=" in the next drop-down menu. Type "2086" in the field box that follows.



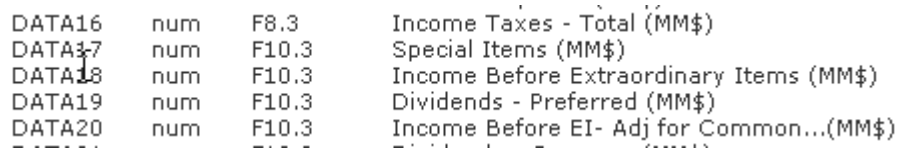
   If you are interested in more information on SIC codes, refer to the links provided in the following page:

   http://www.kellogg.nwu.edu/researchcomputing/sic-codes.htm

   The pages on Compustat and CRSP include a link to that page.

4. In "Step Three: Variables" the variables are organized in three panels. The first two panels contain descriptive information, while financial statement items are in the last panel, in a scroll box. Items are identified by their Compustat mnemonics.

To find a specific item, click on the "Documentation" link for the brief descriptions and write down the data or footnote numbers:



DATA16	num	F8.3	Income Taxes - Total (MM$)
DATA17	num	F10.3	Special Items (MM$)
DATA18	num	F10.3	Income Before Extraordinary Items (MM$)
DATA19	num	F10.3	Dividends - Preferred (MM$)
DATA20	num	F10.3	Income Before EI- Adj for Common...(MM$)

For this exercise, in the first panel, "Identifying Information/Codes", select "Ticker" and "Company Name". On the left side in the third panel, scroll down until you find "DATA12". Click on it, scroll down until you see "DATA18", press the Control (Ctrl) key and click on DATA18. Do the same for "DATA29". If you do not press the Control key before clicking on additional variables, you will loose the previous selections.

5. In "Step Four: Output" select "comma-delimited text (*.csv)" and click on the "Submit Request" to execute your query.

If you are interested in financial statements in their standard format, select the "Tools" link on WRDS toolbar on the top of the screen.

First version: March 23, 2001
Last update: April 8, 2004
-pll

