Title: Reuse of economic data: aggregating and storing vintages of economic data for future use

Authors: Katrina Stierholz, Federal Reserve Bank of St. Louis

Christian Zimmermann, Federal Reserve Bank of St. Louis

Subject: data preservation and reuse in the context of a data repository

The Federal Reserve Bank of St. Louis has a unique aggregator and repository of economic data, FRED (Federal Reserve Economic Data), with over 61,000 series from 49 sources. These economic data have been made available since the early 1990s, initially with a limited focus of providing only the most recent data from the most popular series. Over time, the mission and the capabilities of the FRED repository have grown—first through the addition of more data, then graphing and other end-user tools, and then the collection of "vintage" data. Indeed, one of the most significant features added to FRED was the capability to capture and easily reproduce various vintages of data. We explain more about vintage data below.

Economic data have some interesting features that often do not apply to other data series. Because it is so important that economic data are released in a timely manner, economic data are initially released with incomplete information. For instance, the U.S. Bureau of Labor Statistics's monthly employment report is issued on the first Friday of the following month and is the first report of a month's employment (e.g., on the first Friday in March, February data are released). The establishment portion of the survey, from which the payroll employment numbers are compiled, is a result of the BLS's survey of over 145,000 companies and government agencies, representing about 557,000 individual worksites (United States. Bureau of Labor Statistics).

The information is collected mid-month; however, given the large number of surveys and business priorities, the first preliminary release has only an average 72.5% rate of return. The following month, a revision to the previous month's data is released; because additional survey responses have been received by the BLS, the accuracy of the information improves. Finally, the third month allows for one last revision, by which time the BLS will have received around 94% of the survey responses (United States. Bureau of Labor Statistics.). But this is not the end of the revision process: Every year the data receive "benchmark revisions," which re-anchor estimates to nearly complete employment counts available from the unemployment insurance tax records filed by the State Employment Security agencies (United States. Bureau of Labor Statistics)). Economists and market watchers frequently pay close attention to the employment report, which is often cited as the most important economic data series (McBride) (Baumohl, 2008) and plays an important role in economic policymaking.

The employment situation report is not the only example of data revision. In fact, revisions of economic data are the norm. Major economic data series are revised multiple times over the course of several months, and then again on a yearly or less frequent basis. This can present problems for economists: For many years economists and librarians both have relied on the most current data, as they were seen as the most accurate and reliable. However, over time, it has become clear that important information is lost when only the latest versions of data are stored. Librarians would discard old data releases when the new monthly issue arrived, so users would benefit from the most accurate data. Likewise, economists would work with the revised data, even when attempting to recreate conditions to test a forecasting tool or possibly test other economists' work. In addition, economists used revised data to evaluated policy decisions, despite the fact their information was different from what policymakers had at the time. (Policymakers often focus primarily on current data, which are often the data that are most subject to revision.)

Take the study by John Taylor, one of the most cited in economics, which analyzed the policy actions of the Federal Reserve. Taylor found that the Fed adheres to rules during periods of low inflation, rules that map current inflation and production to policy interest rates. However, these rules were not followed during the high inflation periods of the 1970s (Taylor, 1993). This study was based on historical, revised data. But when Anastasios Orphanides redid the same study with vintage data, he found that the high and low inflation periods were undistinguishable: Policymakers were adhering to the same rules given the information they had at the time (Orphanides, 2002). The vintage of the data made all the difference.

Online provision of data *should* have made archival records easier to compile and access, but it has not. Most agencies issue their press releases, update the data online, and write over the old data, replacing the old file with the new file. The digital storage costs and manual effort required have been perceived as too burdensome, and the return on the investment has not been evident.

To address these issues, the Federal Reserve Bank of St. Louis began to modify its FRED database to capture and preserve the historical, vintage data. In essence, the entire FRED database was rebuilt. What we now call FRED is really a subset—the current version—of the larger database, ALFRED. ALFRED stands for ArchivaL FRED and refers to archived FRED data. This is not readily visible to most data users, because most users simply want access to the current economic data and use FRED exclusively. (In fact, we don't do much to clarify that distinction; we see the two audiences as very different). There are economists, however, for whom these vintage data are very important and for whom ALFRED represents in some cases information that was otherwise unobtainable. In all cases, this information has been, at the very least, time-consuming and painstaking to collect.

The vast majority of data points in ALFRED's original release were captured over many years as part of our multiple back-ups of the FRED database. Sometimes daily and always at least weekly, these backups were the backbone for the construction of the earliest ALFRED data. Additional verification of the data was completed by our research analysts and interns. Most of the data points in ALFRED go back to 1996 or to the date when the data series were first added to FRED. For a very few important series, we went back and found the news releases and newspaper articles and recreated those data points manually. As of 2006, when ALFRED went live, all ALFRED data are captured automatically as part of the data upload for FRED data. We still run quality control checks on the data to verify their accuracy.

ALFRED, as a database, is based on Snodgrass' work (Snodgrass, 1999), which describes a database structure that allows for efficient capture of temporal information and is built to capture data in the most efficient way possible. ALFRED captures the data based on a time interval: For instance, in the Employment Situation Report, the month of March 2012 is a time interval for reporting purposes. For each observation, there are three pieces of information:

- a measurement interval—that is, the time period it applies to (e.g., March 2012);
- a validity interval—that is, the time period it is true for (e.g., from April 5, 2012, until May 2, 2012); and
- a transaction interval—that is, the date the information was entered into the database (to allow for tracking of data entry errors). The transaction interval changes only when there is a change in the data.

The most current observations always have an infinite end date for the validity interval, so for series that are never or rarely revised, the additional observations are very few. And, for nearly all series, after the initial revisions, typically only one revision occurs each year.

Because new observations are added only if there is a revision, the data are stored as efficiently as possible. In other words, if there is no revision to an observation, then all of the other pieces of information are unchanged.

In sum, economists use the data stored in ALFRED to capture information from a point in time; they use that information for their work: perhaps a forecasting model or a reassessment of the work of others, including a replication of another economist's work. As a data repository, ALFRED easy provides access to data from a wide variety of national and international sources that would otherwise be difficult to access. Additional features of ALFRED add to its value, such as user accounts and the ability to send links to series and data sets. Today, ALFRED is the only economic time series database that captures this information and preserves it for future research.

Works Cited

Baumohl, B. (2008). *The Secrets of Economic Indicators: Hidden Clues to Future Economic Trends and Investment Opportunities.* Upper Saddle River, New Jersey: Wharton School Publishing.

McBride, B. (n.d.). *Updated List: Ranking Economic Data*. Retrieved 2 7, 2013, from Calculated Risk: http://www.calculatedriskblog.com/2011/06/updated-list-ranking-economic-data.html

Orphanides, A. (2002). Monetary Policy Rules and the Great Inflation. *American Economic Review*, 115-120.

Snodgrass, R. T. (1999). *Developing Time-Oriented Database Applications in SQL.* San Francisco: Morgan Kaufmann Publishers, Inc.

Taylor, J. (1993). Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy*, 195-214.

United States. Bureau of Labor Statistics. (n.d.). *Current Employment Statistics - CES (National)*. Retrieved 2 7, 2013, from http://www.bls.gov/ces/

United States. Bureau of Labor Statistics). (n.d.). *Frequently Asked Questions (FAQs)*. Retrieved 2 7, 2013, from Current Employment Statistics - CES (National): http://www.bls.gov/ces/cesfaq.htm#revisions

United States. Bureau of Labor Statistics. (n.d.). *CES Registry Receipts by Release*. Retrieved 2 7, 2013, from Current Employment Statistics - CES (National): http://www.bls.gov/web/empsit/cesregrec.htm