Economic Data as Snapshots in Time

*by Katrina Stierholz **

The Federal Reserve Bank of St. Louis has initiated two new projects, FRASER and ALFRED. These two projects share one goal: to provide better access to historical economic data. But the projects differ in their audiences and uses. FRASER is an image archive of economic statistical publications, from government or near-government sources (aka the Fed). ALFRED is a machine-readable archive that allows researchers to pull real-time data in handy formats such as Excel. These two projects build on FRED. FRED, which stands for Federal Reserve Economic Data, is our most heavily-used data product. FRED offers over 3,000 economic and financial time series drawn from government and commercial sources.

REASON FOR PROJECT

Several years ago, Bob Rasche, the Research Director at the St. Louis Fed, began searching for economic data that would tell him exactly what economists knew at a particular point in time. He wanted to be able to create a database that would give him answers to historical real-time data questions. Researchers at the Philadelphia Fed had also been looking at this issue, and they had created the Real-Time Data Set for Macroeconomists, the first real-time datasets for researchers to use. This dataset is useful, but the number of variables is relatively small. The St. Louis Fed seeks to build on that work.

Real-time data represents a point in time, either the moment we are in or sometime in the past. The St. Louis Fed's FRED database is real-time data, but only for the current moment. And while the data in FRED is historical, it does not contain the actual numbers that economists saw in the past—it contains the data that has been revised for the historical time period. That's because many economic time-series are revised (and revised again and again). In order to see the data at a particular point in time, as it was seen by contemporaries of that time, the researcher needs some way of accessing that original data. Unfortunately, those data are much more difficult to obtain.

Economists have seen the need for this data, to answer questions about economic research and economic policy.

1. Replicating other economists' work

FRASER and ALFRED will offer the ability to reproduce other economists' work. While researchers in other fields have been documenting their studies and offering other researchers the chance to replicate their work, economists have been less likely to replicate other economists' work. A significant reason is the difficulty in replicating both the program and the dataset. When economists publish papers, they often do not publish the datasets and the program code used to generate their results. Without this information, it is virtually impossible for other economists to replicate their work and check for errors. [See Dewald 1986.] The St. Louis Fed became a leader in replication when Dewald become research director in 1992. Since that time, for our publication, the *Review*, we require that the data that supports a paper be published (on the web) alongside the article.

2. Sensitivity of Results to Vintages of Data

Another possible use for this historical real-time data is checking the results of studies using a variety of vintages of the same series, all of which contain the preliminary data. An economist could check a model's usefulness by using the real-time data and have the final data to use as a check. It is also useful to examine what economists knew at the time, to see if their policy decisions made sense. That is, based on what we knew then, would this model predict the way things turned out or does this policy decision make sense? [Croushore's August 2004 manuscript details this nicely.] Only real-time data can answer that question.

3. Expectations and policy making with initial data release

Yet another reason for using real-time data involves issues that surround expectations. The information that economists and other policy makers have at the time that decisions are made may change. Sometimes once, sometimes many times. Policy decisions may be made on these "not great, but all we have" numbers – and so it is important to preserve the historical information used for those decisions. Looking back at decisions made, it is important to be aware of what the data *at the time* said, not what they say now. Evaluating the changes in these revisions will also help economists evaluate/properly weight the reliability of the first run numbers.

As an example, consider the numbers published for GDP. GDP is measured quarterly. New measurements, for a variety of quarters, are published each month. Most hotly anticipated, of course, is the data for the most recent quarter. The first release, or measurement, for that quarter is the "advance" release. For the 1st quarter of 2004, this number was released on April 29, 2004. The second published number is known as the "preliminary" estimate — and it was released on May 27, 2004. The third published number is known as "final", and was released on June 25, 2004. Unfortunately, the "final" value isn't final! A revised number often is published the following month (although it lacks a name) and thereafter revisions are published annually. Eventually, the numbers are revised further roughly every five years. [See Croushore 2004.]

In the process of gathering release dates for his own research, Bob Rasche saw the value of this information for other economists. He wanted to build a database that would include several data series, along with the information from each revision, and that would have the dates that the information was good for: think of it as an expiration date for the data. He was very interested in knowing both the data and the date those data were released. Rasche decided to make it available to the rest of the world; first as the publication (a scanned document via FRASER), and then as a database with data points and release dates (ALFRED).

So, for these reasons, Rasche decided to go with a two-pronged approach to preserving this information and making it widely available. ALFRED and FRASER have been conceived as a pair, but they are very different data products, and will, in the end, serve different audiences. Both build on the original FRED database concept.

FRED

The Federal Reserve Bank of St. Louis has FRED (Federal Reserve Economic Data), which offers over 3,000 economic time series, including banking, financial, employment, monetary, interest rate data, and more. The data is collected from a variety of sources — the Board of Governors and the US Government, as well as some commercial sources. It is presented in useful ASCII or Excel formats, and the data can be downloaded in large data sets or as a single item.

FRASER

FRASER — our Internet image library — is a natural outgrowth of gathering all of that lost information from the press releases. As we discovered how difficult it was to find press releases, it became clear that information was being lost. And, if it hadn't disappeared yet, it would soon, as libraries hurry to clean off their shelves. The easiest method for providing this information is to scan serial economic publications that have a variety of economic data. So we began scanning all those press releases for various economic time series. We then spent weeks and months hunting down all the gaps in our information, taking documents from every librarian that would let us beg and borrow their documents.

FRASER, as an image archive, includes tools that allow for sophisticated retrieval of statistical tables linked over many years of publication. We started with the monthly *Economic Indicators*, a publication of the Joint Economic Committee, and we added some Federal Reserve and Federal Government titles to the mix—things like *Banking and Monetary Statistics*, *All Bank Statistics*, and *Business Statistics*. These are basic titles, with lots of economic data. Much of this data will not be available on ALFRED anytime soon; there's just too much to enter, and it would require so much work. But the data are available for anyone who wants it.

FRASER as part of the GPO digitization project

FRASER will give historians access to what policy-makers and economists (and the public, for that matter) knew at the time. It will also be a part of the national bibliography of government publications that the GPO is coordinating in our natural niche of economic statistical publications. We have scanned the material in the manner requested by GPO. FRASER fulfills the need for wide access to historical economic data, at a relatively low cost. We have added more publications to our list of scanned periodicals: the *Business Conditions Digest*, *Survey of Current Business*, the *Economic Report of the President*, and the *H.6 Money Stock Measures* (a release from the Board of Governors); these will be posted soon.

To make the data in these publications more accessible, we've done a couple of things. One, we've added keywords to the metadata so that searching is useful. Another is that we've linked all of the tables within a publication, so that if you find a table that answers your research need, you can download the same table over time. We've OCR'd the documents, so the text is searchable for terms (for example, "gold"). We have added title continuation information and SuDoc numbers to help users. And, finally, we've made it all searchable via the Autonomy search engine. This combination gives users all the access that I can imagine they need, except for one thing. This combination gives users all the access that I can imagine they need, except for one thing. While we have OCR'd everything, including the data, we do not allow users to extract or copy the text or numbers, because the OCR has not been verified. Users can save the PDF, or print it out, but the OCR has been locked so that it isn't accessible. Not that we want to deny users this but, because we haven't corrected the OCR, we're concerned about naïve users who might copy the information into a spreadsheet and not realize that some numbers are incorrect. It is particularly difficult to spot errors in uncorrected OCR tables. Correcting the OCR is time

consuming and tedious (which translates to expensive). If the use of the material is high enough to warrant OCR correction, we'll put it into ALFRED, because ALFRED is our database of real-time information that has been verified. The user will have more access. The user will have more access, in a better format, than if we leave it in FRASER. Any really intense user can easily OCR the PDF files themselves, using a variety of readily available commercial packages — in which case the user is responsible for all errors, not us.

FRASER has been built on the JSTOR model in many ways. We have scanned each publication at 600dpi in TIFF format, and then digitally "cleaned" it of non-printed marks (such as handwritten notations, creases, or stapler marks) or imperfections created during the scanning process. The clean file is then run through an optical character recognition (OCR) software application that images the file, introduces metadata, and converts the file to Adobe Portable Document Format (PDF), as well as compressing the scanned image from 600 dpi to 300 dpi.

We have complied with the United States Government Printing Office requirements for scanning, because if and when the day comes that we need to migrate this material (should PDF become an obsolete format), we want to be in a very common format, so that there is a common solution. If we did something fabulous but unique, we might not have good options for migrating the information.

FRASER is a low-cost way of providing a large amount of statistical data, and it allows for uses of the data in ways that we have not imagined. FRASER is useful to a wide variety of audiences—not just economists but historians who want to see a contemporaneous look at economic data. FRASER recognizes that the largest cost in many research projects is locating the published data; entering the data into the computer is the easier part.

The second project we are undertaking is called ALFRED. It builds an archival feature onto our foundational FRED database, but with less data than is available on FRASER.

ALFRED (Archival FRED) data

ALFRED will provide sophisticated data for economic researchers in a much more useable format than FRASER. The data will all be available in text and Excel spreadsheets, and will allow the user to pull multiple versions of the same data set, using different points in time as a reference.

ALFRED will be populated initially with the archived FRED data. At first, the data will go back to December of 1996, as we have taken a snapshot of our data every Friday of each month since 1996. It is this data, along with the information compiled about the release dates, that will populate ALFRED. The release dates for economic data have become standardized over the years, and holiday information is easier to get. So verifying release dates for the fairly recent data has been relatively straightforward. The first release of ALFRED will contain the data from these snapshots and the release dates. Because the data are being verified, the first release will contain the information for employment, CPI, and PPI; not for every single series in FRED. A user will be able to download the data for any one of these series for many different dates. For example, she could download the employment numbers available at several different points in time. So, if she wanted to know what employment numbers were available before every FOMC (Federal Open Market Committee) meeting, where they target the Federal Funds Rate, she could download the employment numbers that were available each time they met... and see exactly what those policymakers saw when they were making policy. Not the revised numbers that came out later, but the information they actually had.

Bob Rasche has also collected the release dates and data for 25 data series going back before December 1996. These have been entered by hand (by an intern, who probably regrets ever taking that position). These numbers were then verified by a research analyst at the Bank. These will also be available in the next release of ALFRED. Later, more historical data will be included, and we plan to provide links between FRASER and ALFRED, so that if ALFRED doesn't have the information in its database, the information can be retrieved from a document stored on FRASER.

Our initial release of ALFRED to customers on the Internet will occur in July 2005 as a feature within FRED. ALFRED will work in two steps. First, a user locates a needed data series in FRED; then, second, s/he has the option of getting the most recent data and/or additional historical data. The user can select the range of desired historical data.

To our customers, ALFRED will look very much like FRED, but with additional available data. Internally, however, the FRED database has been reconstructed in an entirely new way, in order to make this project scalable. The database was created by George Essig, a senior web developer at the St. Louis Fed, and he devised an ingenious way to store the data so it wouldn't take up a ton of server space, and so that it would contain all the information necessary for researchers. George has used a concept found in Richard Snodgrass' publication Developing Time-Oriented Database Applications in SQL (2000) that suggests a sophisticated way to store the data.

In addition to the number that represents the observation for that moment, each data point also has two separate, additional pieces of information about it. The first is a measurement of the time interval that the data covers (for instance, if the data covers March of 2005, it would have

a time interval measurement of 3/1/2005 to 3/31/2005). Then, it has a second bit of information: the dates for which this information was valid. So for the first release of the March 2005 employment number, the period of validity would be from April 1, 2005 (the date of the first release) until May 6, 2005 (the date of the second release). It would have an open-ended date until the data was updated, so if it is never updated, the number stays valid forever. However, if the data changes, the program caps off the end date of the expired number and adds the new number. This feature keeps the back-end of the program relatively simple, and the total file size relatively small.

For more details on what George has done to create the database, and the time interval measurement, validity intervals, and transaction intervals, please read a paper written by one of our economists, Richard Anderson: "Replicability, Real-Time Data, and the Science of Economic Research", March 2005. Dick Anderson has several papers on this topic.

FRED and ALFRED are not two separate databases, but one database. Absent historical data, FRED would have 1.1 M data points; in the new combined FRED+ALFRED, the database has 2.2 M data points.

ALFRED will start small, and live within FRED for a while. The first version will allow users to locate a dataset, and then choose the date they want. We anticipate that ALFRED will serve economic researchers who are looking at modeling their work using real-time data from a variety of vintages, and also economists who are interested in replicating other economists' work. FRED is used by a large number of people with a wide variety of purposes. By comparison, ALFRED will reach a small audience with a narrow focus. Also, while ALFRED will have excellent retrieval methods, useful for gathering multiple vintages of a series in a single shot, creating the data sets in ALFRED is a labor-intensive and time-intensive project. This will limit both the number of data sets that will go back before December 1996, and the time period that they will go back to. There may also be some series for which we are unable to gather the information before a certain date.

LIBRARY/LIBRARIAN ISSUES

There are also some important library issues that arose during the work on this project. It was sometimes very difficult to find the release date for data. Virtually every library threw out the oldest press releases. In many cases, we found only a few libraries that owned the oldest press releases. Even the issuing agency and Library of Congress did not have these documents.

As well as losing the print press releases, or other "current news" that was subsequently revised, the electronic versions were also often lost. Because disk space was scarce, files were replaced rather than versions added, a loss that is permanent. While we would sometimes find the paper press releases in libraries as a result of benign neglect, in the case of the electronic files there might only be the most recent copy (if that), as all the others had been overwritten.

As data librarians, consider the potential use of snapshots of your data, particularly if your data are subject to revision. I am sure that there are other fields where these issues might apply, and even if you aren't sure if it will apply, consider taking snapshots if your data changes over time.

For many researchers, it important to analyze data as it was available at the time, rather than the perfected data released much later. Important policy decisions are made with this imperfect data. It is important to see the data using the same imperfect lens as economists had at the time it was released; and determining the problems of that data as well as its usefulness is key to developing good models. This data may also help economists create better models, or provide information on the usefulness of other economists' models. We hope that by providing FRASER and ALFRED, the St. Louis Fed will be able to contribute to that work by economists.

References

Anderson, Richard G. "Replicability, Real-Time Data, and the Science of Economic Research" Manuscript, Federal Reserve Bank of St. Louis, March 2005. (Forthcoming, Federal Reserve Bank of St. Louis Review.)

Anderson, Richard G., William H. Greene, Bruce D. McCullough, and H. D. Vinod, "The Role of Data & Program Code Archives in the Future of Economic Research" Federal Reserve Bank of St. Louis, Working Paper 2005-14.

Croushore, Dean. "Forecasting with Real-Time Macroeconomic Data" Manuscript, University of Richmond, August 2004.

Croushore, Dean, and Tom Stark. "A Real-Time Data Set for Macroeconomists" Journal of Econometrics 105 (November 2001), pp.111-130.

Croushore, Dean, and Tom Stark. "A Funny thing Happened on the Way to the Data Bank: A Real-Time Data Set for Macroeconomists" Federal Reserve Bank of Philadelphia, Business Review (Sept./Oct. 2000), pp. 15-27.

Dewald, William G., Jerry Thursby, and Richard Anderson. "Replication and Scientific Standards in Empirical Economics: Evidence from the JMCB Project" American Economic Review, September 1986, 76:4, pp. 1255-57.

Rasche, Robert, Katrina Stierholz, Robert Suriano, and Julie Knoll. "As It Happened: Economic Data and Publications as Snapshots in Time" Presentation at the Federal Depository Library Conference, October 19, 2004, Washington, D.C.

Snodgrass, Richard T. Developing Time-Oriented Database Applications in SQL. Morgan Kaufman, San Francisco. 2000.

Endnotes

[1] Contact: Katrina Stierholz, Federal Reserve Bank of St. Louis, PO Box 442, St Louis MO 63166 314-444-8552, Katrina.L.Stierholz@stls.frb.org

The author is grateful for the comments of Robert Rasche, Richard Anderson, and George Essig.

The views expressed are those of the author and do not necessarily reflect the official positions of the Federal Reserve Bank of St. Louis, the Federal Reserve System, or the Board of Governors.

* The paper was presented at the IASSIST 2005 conference in Edinburgh.