



Replicability, Real-Time Data, and the Science of Economic Research: FRED, ALFRED, and VDC

Richard G. Anderson

This article discusses the linkages between two recent themes in economic research: “real time” data and replication. These two themes share many of the same ideas, specifically, that scientific research itself has a time dimension. In research using real-time data, this time dimension is the date on which particular observations, or pieces of data, became available. In work with replication, it is the date on which a study (and its results) became available to other researchers and/or was published. Recognition of both dimensions of scientific research is important. A project at the Federal Reserve Bank of St. Louis to place large amounts of historical data on the Internet holds promise to unify these two themes.

Federal Reserve Bank of St. Louis *Review*, January/February 2006, 88(1), pp. 81-93.

REPLICATION AND REAL-TIME ECONOMETRICS

During the past 25 years, two themes have flowed steadily, albeit often quietly, through economic research: “real time” data and replication. In *replication* studies, the issue is determining which data were used and whether the author performed the calculations as described; in *real-time* data studies, the issue is determining the robustness of the study’s findings to data revisions. These themes share the same core idea: that scientific research has an inherent time dimension. In both real-time data and replication studies, the time dimension is the date on which particular observations, or pieces of data, became available to researchers. Projects at Harvard University and at the Federal Reserve Bank of St. Louis promise to improve the quality of empirical economic research by unifying these themes.¹

Although replication studies focus on the *correctness* of results and real-time studies on

their *robustness*, economic theory suggests that these are related—the likelihood that an author’s error will become visible to other researchers is an inverse function of the cost of conducting tests for replicability and robustness. Yet, for the profession, excessive emphasis on the criminal-detection aspects of replication (Did the author fake the results? Or did the author cease experimenting prematurely when a favorable result appeared?) has tended to increase the reluctance of researchers to share data and program code. That is, to the extent that the profession overemphasizes the manhunt of David Dodge’s 1952 *To Catch a Thief*, it risks foregoing the benefits of Sir Isaac Newton’s 1676 dictum, “If I have seen further it is by standing on the shoulders of giants.”

The incentives and disincentives for a researcher to share data have been discussed by numerous authors (e.g., Fienberg, Martin, and Starf,

¹ The fallacy that neither real-time data nor replication studies are needed because “important” results always will sift to the top through repeated studies is addressed at length in Anderson et al. (2005).

Richard G. Anderson is a vice president and economist at the Federal Reserve Bank of St. Louis. This is a revised version of paper prepared for the American Economic Association meeting, Philadelphia, PA, January 2005. The author thanks Bruce McCullough of Drexel University, Gary King of Harvard University, and Dan Newlon of the National Science Foundation for their comments. Giang Ho provided research assistance.

© 2006, The Federal Reserve Bank of St. Louis. Articles may be reprinted, reproduced, published, distributed, displayed, and transmitted in their entirety if copyright notice, author name(s), and full citation are included. Abstracts, synopses, and other derivative works may be made only with prior written permission of the Federal Reserve Bank of St. Louis.

1985; Boruch and Cordray, 1985; Dewald, Thursby, and Anderson, 1986; Feigenbaum and Levy, 1993; Anderson and Dewald, 1994; Bornstein, 1991; Bailer, 2003).² Researchers receive a stream of rewards for the new knowledge contained in a published article, which begins with publication and eventually tapers to near zero. Furnishing the data to other researchers invites the risk that a replication will demonstrate the article's results to be false, an event which immediately ends the reward stream. If the replication further uncovers malicious or unprofessional behavior (such as fraud or other unethical conduct), "negative rewards" flow to the researcher.

Creating original research manuscripts for professional journals is craft work. Although often referred to as "knowledge workers," researchers might equally well be regarded as artisans, with creative tasks that include collecting data, writing code for statistical analysis or model simulation, and authoring the final manuscript.³ Similar to the work of other craftsmen, researchers' output contains intellectual property—not only the final manuscript, but also the data and programs developed during its creation. Yet, for academic-type researchers, some of the intellectual property must be relinquished so the work can be published in peer-reviewed journals. This conflict creates a strategic game in which the researcher feels compelled to reveal a sufficient amount of his material to elicit publication, while simultaneously seeking to retain for himself as much of the intellectual property as possible. There are few, if any, models of this process in the economics literature. One such analysis is presented by Anderson et al. (2005), based on the Crawford and Sobel (1982) model of strategic information withholding. A complete presentation of their theoretical analysis is beyond the scope of this paper. The results

buttress, however, the commonsense intuition that so long as withholding data and program code does not reduce the post-publication stream of rewards (and disclosure of data and program code does not increase it), researchers will rationally choose not to disclose data and programs.⁴ Such models largely explain the well-known proclivity of academic researchers in many disciplines, including economics, to keep secret their data and programs. For the progress of scientific economic research, such an equilibrium is suboptimal.

One solution to suboptimal equilibria is collective action. One collective action is for professional journals to archive data and program code.⁵ Such archives—which permit low-cost, anonymous, ad hoc replication—can improve the quality of published research by way of an effect reminiscent of Baumol-like credible threats of market entry. This process was well-described by the University of Chicago's John Bailer (2003) at a recent National Research Council conference:

Of all the public myths about how science is done, one of the broadest and most persistent is that scientific method rests on replication of critical observations [i.e., results]. Straight replication is in fact uncommon, largely, I believe, because no scientist gets much professional credit for straightforward replication unless the findings are critical, there is suspicion of fraud, or there is some other unusual condition such that slavish replication of the methods reported might have some meaning not attached to the first round. Here I exclude replication by an independent investigator for the sole purpose of assuring himself or herself that the original results are correct and that the methods are working properly, as a preliminary to going further in some way.

Overall, replication...seems to be one of those ideals that get a fair amount of discussion

² Some data cannot be shared. Examples include confidential banking data held by the Federal Reserve; micro data held by the Bureau of the Census; and various financial data, including that licensed by the University of Chicago's Center for Research in Security Prices (CRSP). In some cases, the owners/licensors of such data have archived the datasets built and used by individual researchers and made the datasets available to subscribers.

³ Indeed, economists and other scientists often refer to "polishing" a final manuscript, in the spirit of woodworkers or stone masons polishing their work.

⁴ The model of Feigenbaum and Levy (1993), in which rewards to researchers are driven by citations, also suggests that the divergence between the search for truth and rational individual choice will be largest for younger researchers (such as those without academic tenure), who will be less inclined to search for errors than older researchers and less inclined to devote scarce time to documenting their work.

⁵ Historical data, cataloged and indexed by the day on which the data became available to the public, often are referred to as "vintage" data.

but have little influence on behavior. *Perhaps what is most important is that the original investigators publish background and methods with enough detail and precision for a knowledgeable reader to replicate the study if he had the resources and inclination to do so.* [emphasis added]

In a recent article, Pesaran and Timmermann (2005, p. 221) offer a formal statement of the correspondence between the universe of all possible datasets and an article's specific dataset:

Let χ denote the time-invariant universe of all possible prediction variables that could be considered in the econometric model, while N_{xt} is the number of regressors available at time t so $X^t = (x_{1t}, \dots, x_{N_{xt}t}) \subseteq \chi$. N_{xt} is likely to grow at a faster rate than the sample size, T . At some point there will therefore be more regressors than time-series observations. However, most new variables will represent different measurements of a finite number of underlying economic factors such as output/activity, inflation and interest rates.

Below, we use the notation X^t to denote the set of all observations [values, measurements], on a fixed list of economic variables, that have been published as of (up to and including) date t . Assuming that an author has not falsified or erroneously transcribed data values, the true dataset for a published article will be contained within the universe of all such datasets X^t , where t is no greater than the date on which the original author completed his research. Unfortunately, such datasets often are too large to be compiled by individual researchers.

Historical data, cataloged and indexed by the day on which the data became available to the public, are referred to as “vintage” data. Collections of such data— X^t in the notation above—are referred to as “real-time” datasets and are indexed by the date of the most recent data included, t . The first large-scale project to collect and make available to the public vintage macroeconomic data was started in 1991 by the Federal Reserve Bank of Philadelphia to assess the accuracy of forecasts collected in the Survey of Professional Forecasters (Croushore and Stark, 2001). That project, and its data, is referred to as the Real Time

Dataset for Macroeconomists (RTDSM). The design of the Philadelphia RTDSM project seeks to provide snapshots of the values of certain macroeconomic variables as they were available to the public at the end of the 15th day of the center month of each quarter. Hence, although both monthly and quarterly data are included, the dataset's primary value is macroeconomic modeling and forecasting at quarterly frequencies.⁶

Data Projects

This essay discusses two on-going projects to support the collection and dissemination of vintage data. The first, ALFREDTM, is the Federal Reserve Bank of St. Louis's Archival Federal Reserve Economic Data. (Some features of ALFRED are still under development and may not be available at the time this article is printed; see alfred.stlouisfed.org for an introduction, instructions, and updates.) The second, VDC, is Harvard University's Virtual Data Center project. The projects differ significantly from each other, and from the Philadelphia project, in several aspects:

- Data frequency: Both the RTDSM and ALFRED projects focus on macroeconomic data. The RTDSM dataset is designed for a quarterly observational data frequency. The ALFRED project is designed for data at a daily frequency or lower (e.g., weekly, biweekly, monthly, etc)—that is, any data frequency currently supported in the Federal Reserve Bank of St. Louis's FRED[®] (Federal Reserve Economic Data) database. The VDC project, discussed further below, focuses on archiving and sharing complete datasets from specific research studies and articles. As such, it is data-frequency independent.
- Data vintages: The RTDSM project provides snapshots of the values of certain macroeconomic variables as they were known by the public at the close of business on the 15th day of each quarter's center month.

⁶ See the Federal Reserve Bank of Philadelphia's web site for details and documentation, available at www.phil.frb.org/econ/forecast/realindex.html, as of October 18, 2005.

The ALFRED project, operating at a daily frequency, provides daily (end-of-day) snapshots of the values of all variables in the FRED database. The VDC project, because it stores complete datasets as provided by researchers, has no explicit vintage component. The lack of a vintage component restricts the value of its design, for economists, to replication alone—an important function, but distinctly different from studies of robustness that require vintage-indexed data such as RTDSM and ALFRED.

- Data updating: Neither the RTDSM nor VDC projects have a mechanism to automatically update their data vintages. Data are added to RTDSM as Philadelphia staff determine which figures were available to the public on the specified days. Datasets are added to the VDC project as researchers place them on the Internet and the VDC servers index their location. The architecture of ALFRED differs. “Under the hood,” ALFRED and FRED share the same database architecture. In this shared design, data values on FRED that are revised—that is, replaced with newly released numbers—are automatically added to ALFRED as vintage data. Combined with a history of release dates for major economic indicators such as GDP and employment, ALFRED uniquely provides a day-by-day vintage snapshot of the evolution of macroeconomic variables. This architecture, as discussed further below, uniquely allows ALFRED to be used for both replication and robustness studies.

Further discussion of ALFRED’s time-indexed architecture is contained in a subsequent section.

The Literature

Studies that explore the sensitivity of empirical results to data vintage and how policymaking might incorporate the data revision process have a long history. Early studies are reviewed by Croushore and Stark (2001). Selected recent studies include Neely, Roy, and Whiteman (2001), Orphanides (2001), Stark and Croushore (2002), Christoffersen, Ghysels, and Swanson (2002),

Orphanides and van Norden (2002, 2003), Bernanke and Boivin (2003), Faust, Rogers and Wright (2003), Koenig, Dolmas and Piger (2003), Svensson and Woodford (2003, 2004), Clark and Kozicki (2004), and Kishor and Koenig (2005). The analysis of vintage, or “real-time,” data also is a popular conference topic—examples include the Federal Reserve Bank of Philadelphia’s 2001 “Conference on Real Time Data Analysis,”⁷ the Bundesbank’s 2004 conference “Real-Time Data and Monetary Policy,”⁸ and the CIRANO/Bank of Canada’s October 2005 “Workshop on Macroeconomic Forecasting, Analysis and Policy with Data Revision.”⁹

One of the earlier experiments to demonstrate the dependence of empirical results on data vintage is reported in Dewald, Thursby, and Anderson (1986). During their project at the *Journal of Money, Credit, and Banking* from 1982 to 1984, a large number of authors, when asked to submit datasets and programs, replied that they did not save publicly available macroeconomic data because the data could easily be collected from published sources and their empirical results were nearly invariant to the vintage of the data.¹⁰ To test this assertion, Dewald, Thursby, and Anderson examined in detail one article, Goldberg and Saunders (1981), which contains a model of the growth of foreign banks in the United States. The article’s authors furnished their banking data (which had required considerable effort to collect) but not their macroeconomic data, which they said had been collected from various issues of the *Survey of Current Business* and *Federal Reserve Bulletin* with no record made of which numbers were obtained from which issues. Dewald, Thursby, and Anderson collected from the *Survey of Current Business* all published values on three macroeconomic variables used in the article

⁷ Program and papers were available at www.phil.frb.org/econ/conf/rtdaconf.html, as of October 18, 2005.

⁸ Program and papers were available at www.bundesbank.de/vfz/vfz_daten.en.php, as of October 18, 2005.

⁹ Program and papers were available at www.cirano.qc.ca/financegroup/Real-timeData/program.php, as of October 18, 2005.

¹⁰ Even authors who submitted data seldom noted the dates on which they collected the data or the dates on which the data had been published.

(imports, investment, and GNP) during the period 1972:Q4 through 1982:Q3.¹¹ From these data, they estimated 500 variants of the Goldberg-Saunders model, summarizing the results in a set of histograms (Dewald, Thursby, and Anderson, 1986, p. 599). Overall, the coefficient estimates obtained varied widely and the modal values often were far from the coefficients in the Goldberg and Saunders article.

DATA SHARING

The arguments above suggest that the quality of empirical economic science is positively correlated with the extent to which researchers preserve and share datasets and program code. This theme is commonplace in science. In 1979, the Committee on National Statistics of the National Research Council (of the National Academy of Sciences) sponsored a conference on the role of data sharing in social science research. A subsequent subcommittee on sharing research data stated the issues clearly (Fienberg, Martin, and Starf, 1985, pp. 3-4):

Data are the building blocks of empirical research, whether in the behavioral, social, biological, or physical sciences. To understand fully and extend the work of others, researchers often require access to data on which that work is based. Yet many members of the scientific community are reluctant or unwilling to share their data even after publication of analyses of them. Sometimes this unwillingness results from the conditions under which data were gathered; sometimes it results from a desire to carry out further analyses before others do; and sometimes it results from the anticipated costs, in time or money, or both.

The Committee on National Statistics believes that sharing scientific data with colleagues reinforces the practice of open scientific inquiry. Cognizant of the often substantial costs to the original investigator for sharing data, the committee seeks to foster attitudes and practices within the scientific

community that encourage researchers to share data with others as much as feasible.

The subcommittee offered 16 recommendations (see the appendix) for improving the quality of social science research through data sharing. The recommendations are so straightforward as to seem self-evident: Sharing data should be standard practice; researchers should retain data for a reasonable period after publication; researchers requesting data should bear the costs of providing data; funding organizations should encourage data sharing by requesting a data-sharing plan in requests for funding; and journals should encourage authors to share data. Yet, two decades later, most of these are not yet standard operating procedures in economic research.

At the time of the National Research Council's 1979 conference, the National Science Foundation's (NSF) policies embodied many of the Council's later recommendations. The NSF Grant Policy Manual NSF-77-47, as revised October 1979, states the following in paragraph 754.2:

Data banks and software, produced with the assistance of NSF grants, having utility to others in addition to the grantee, shall be made available to users, at no cost to the grantee, by publication or, on request, by duplication or loan for reproduction by others...Any out of pocket expenses incurred by the grantee in providing information to third parties may be charged to the third party.

Subsequent to publication of Dewald, Thursby, and Anderson (1986), the NSF's social science program adopted a policy of requiring that investigators place data and software in a public archive after their award expired.¹² The NSF also began asking researchers, in applications for subsequent funding, what data and software from previous awards had been disseminated. Today, the NSF policy is clear. The NSF's current *Grant Proposal Guide* (NSF 04-23, effective September 2004), section VI, paragraph I, states that

NSF advocates and encourages open scientific communication...It expects PIs [principal

¹¹ The published Goldberg-Saunders article used data through 1980:Q1; Dewald, Thursby, and Anderson collected data through the December 1982 issue of the *Survey*.

¹² I am indebted to Dan Newlon, head of the economics program at the National Science Foundation, for the information contained in this paragraph.

investigators] to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work. It also encourages grantees to share software and inventions, once appropriate protection for them has been secured, and otherwise act to make the innovations they embody widely useful and usable. NSF program management will implement these policies, in ways appropriate to field and circumstances, through the proposal review process; through award negotiations and conditions; and through appropriate support and incentives for data cleanup, documentation, dissemination, storage and the like. Adjustments and, where essential, exceptions may be allowed to safeguard the rights of individuals and subjects, the validity of results and the integrity of collections, or to accommodate legitimate interests of investigators.

The NSF's *Grant Policy Manual* (NSF 02-151, effective August 2, 2002), paragraph 734, "Dissemination and Sharing of Research Results," contains similar statements:

b. Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. Privileged or confidential information should be released only in a form that protects the privacy of individuals and subjects involved. General adjustments and, where essential, exceptions to this sharing expectation may be specified by the funding NSF Program or Division for a particular field or discipline to safeguard the rights of individuals and subjects, the validity of results, or the integrity of collections or to accommodate the legitimate interest of investigators. A grantee or investigator also may request a particular adjustment or exception from the cognizant NSF Program Officer.

c. Investigators and grantees are encouraged to share software and inventions created under the grant or otherwise make them or their products widely available and usable.

d. NSF normally allows grantees to retain principal legal rights to intellectual property developed under NSF grants to provide incentives for development and dissemination of inventions, software and publications that can enhance their usefulness, accessibility and upkeep. Such incentives do not, however, reduce the responsibility that investigators and organizations have as members of the scientific and engineering community, to make results, data and collections available to other researchers.

With such a strong policy in place, data warehousing in economic research should be commonplace. In fact, it remains rare. As of this writing, 10 professional economics journals have data and/or program archives: *American Economic Review*; *Econometrica*; *Macroeconomic Dynamics*; *Journal of Money, Credit, and Banking*; *Federal Reserve Bank of St. Louis Review*; *Economic Journal*; *Journal of Applied Econometrics*; *Review of Economic Studies*; *Journal of Political Economy*; and *Journal of Business and Economic Statistics*. Some require both data and program files, others only data. In addition, a public archive for data and programs from published articles in any professional journal has been maintained since 1995 by the Inter-university Consortium for Political and Social Research at the University of Michigan; except for articles related to the Panel Study of Income Dynamics at Michigan, all of the economics-related articles' data and programs in the archive are from the Federal Reserve Bank of St. Louis *Review*.

The ALFRED project has the potential, for macroeconomic research, to eliminate the need for journals to store authors' datasets. ALFRED, as explained further below, will sharply reduce the costs that a researcher in macroeconomics would bear in documenting, storing, and distributing the data used in a research project. In this aspect, data archives at journals are a "complementary technology" to the vintage archive structure of ALFRED, and, both in concept and execution, more similar to the dataset-archiving-and-indexing design of the VDC project.

DATA WAREHOUSING: THE FRASER AND ALFRED PROJECTS

The collection and distribution of data has classic public-good characteristics, including economies of scale, network effects, and first-mover advantages. Yet, large-scale systems for archiving and distributing economic data are rare.

As noted above, data warehousing for “real time” economic research began with the Federal Reserve Bank of Philadelphia (Croushore and Stark, 2001, p.112):

In creating our real-time data set, our goal is to provide a basic foundation for research on issues related to data revision by allowing researchers to use a standard data set, rather than collecting real-time data themselves for every different study.

Two current data-warehousing projects of the Federal Reserve Bank of St. Louis are in the same spirit: FRASER® (Federal Reserve Archival System for Economic Research) and ALFRED. Together, these projects seek to provide a comprehensive archive of economic statistical publications and data. Initially, the projects will focus on government macroeconomic data but eventually will be expanded to include less aggregate data.

FRASER

The FRASER project (fraser.stlouisfed.org) is an Internet archive of images of statistical publications. The long-term goal, essentially, is to include all the statistical documents ever published by the U.S. government, plus other selected documents from both private and public sources. FRASER is an “open standards” project—that is, any organization that wishes to submit images is encouraged to do so provided that the images satisfy the requirements suggested by the U.S. Government Printing Office’s committee of experts on digital preservation.¹³ To date, however, most contributions have been boxes of printed paper materials, rather than images.

¹³ U.S. Government Printing Office (2004); Federal Reserve Bank of St. Louis (2004).

ALFRED

The ALFRED project is an archive of machine-readable real-time data. Since 1989, the Federal Reserve Bank of St. Louis has provided data to the public on its FRED system.¹⁴ Initially, ALFRED will be populated with archived FRED data beginning December 1996. Later, other historical data will be added, including data extracted from FRASER images. Links will be provided between ALFRED data and the historical FRASER publications. The purpose of the FRED data system is to distribute the most-recent value for each variable and date; the purpose of the ALFRED system is to distribute in a similar method all data values previously entered into FRED, plus additional data. (Again, ALFRED is a work under construction; some features outlined here are part of the ALFRED architecture, and some are not yet available.)

ALFRED is a relational database built on PostgreSQL. Users are able to request subsets of data by means of an automated interface. In ALFRED, every data point is tagged with the name of its source and its publication date. For real-time projects, a researcher need only submit a variable list and a desired range of vintages (that is, “as of” dates); ALFRED will return all values for those variables that were available during the specified date range, each tagged with its as-of date (that is, its publication date). For replication studies, in the unlikely circumstance that the original researcher used the most recently published values for all variables and dates, a researcher need submit only a list of variable names and the as-of date when the original researcher collected his data. In the more common circumstance, in which the original researcher is uncertain whether he collected the then-most-recent data, a putative range of collection (as-of) dates may be submitted; with luck, some mixture of the retrieved values perhaps will reproduce the original published results.

Combined, the FRASER and ALFRED projects are a “statistical time machine” that, on request, furnishes to researchers both universes of data, χ , and time-indexed “real-time” subsets, N_{xt} .

¹⁴ Initially, FRED operated as a dial-up computer bulletin board system. In 1995, shortly after the release of version 1.0 of the Mosaic web browser, FRED appeared as a web site on the Internet.

As of this writing (October 2005), portions of ALFRED remain in development and not all features are implemented. Essential, however, will be a scheme to uniquely identify the historiography of data retrieved from ALFRED. The current design proposal includes the concept of a *research dataset signature*, or RDS. The proposed RDS is a human-readable string of ASCII characters that uniquely identifies a data series extracted from FRED or ALFRED. A dataset containing multiple time series (variables) will have an RDS for each series. The proposed character encoding pattern is as follows:

- 1-20: the FRED/ALFRED variable name¹⁵;
- 21-22: the frequency code, e.g., “A”, “Q”, “BW”;
- 23-25: the seasonal adjustment code, either the string “SA” or “NSA”¹⁶;
- 26-34: the 24-hour Greenwich Mean Time date on which the data were downloaded, in the form DDMMYYYY (that is, the form “27JAN2001”);
- 35-42: the 24-hour Greenwich Mean Time time-of-day at which the data were downloaded, in the form HH:MM:SS (HH=hour, MM=minute, SS=second).

Greenwich Mean Time is used because it does not vary with the geographic location of the researcher or the season of the year.¹⁷ These date and time-of-day formats are sufficiently general to accommodate researchers located anywhere in the world. Within each RDS string, character fields will be left-justified and padded on the right with spaces (ASCII 32). The RDS is somewhat shorter than might be anticipated because it does not include a “start” and “end” date. In FRED

and ALFRED, users are not permitted to select/download a subset of a time series. A time series must be downloaded in its entirety or not at all. (The user is free to discard any unwanted data after download.) This permits a shorter signature.¹⁸ Finally, as plain text, RDS strings may easily be included in working papers and journal articles.

The proposed architecture for ALFRED follows, in part, the bitemporal SQL (structured query language) database structure of Snodgrass and Jensen (1999).¹⁹ In this design, each datum (observation) for each time series will be stored with three 2-element date vectors. One vector demarcates the beginning and end of the *measurement* interval for the observation, the second the beginning and end of the *validity* interval, and the third the beginning and end of the *transaction* interval.

Measurement intervals are straightforward. The measurement interval for GDP during 2004:Q1, for example, would be {1Jan2004, 31March2004}; a daily interest rate might have an interval of the form {5Jan2004, 5Jan2004}; and a monthly average interest rate might have an interval of the form {1Jan2004, 31Jan2004}. This system encompasses, in a uniform way, all data frequencies.

Validity intervals demarcate the time periods during which a datum was the most recently published value. As an example, consider 2004:Q1. See Table 1. During 2004, the Bureau of Economic Analysis published four measurements on 2004:Q1 GDP: April 29 (“advance”), May 27 (“preliminary”), June 25 (“final”), and July 30 (this fourth value has no commonly used label). During 2005, the BEA published one measurement, on July 29, as part of the 2005 benchmark revisions. The validity intervals shown reflect these dates. Note that the *fifth* validity interval is open-ended and will remain so until the next revised value is published.

¹⁵ Currently, all FRED/ALFRED variable names are 8 characters in length and composed only of the characters 0 to 9 and A to Z—that is, ASCII characters 48 to 57 and 65 to 90. The signature’s name field, to allow future expansion, is 20 characters in length and allows the underscore, ASCII 95, as well as 0 to 9 and A to Z.

¹⁶ In the current FRED nomenclature, the seasonally adjusted character of the series can be inferred from the variable name. This field is included to increase the human usability of the signature and for possible future expansion of the FRED/ALFRED nomenclature.

¹⁷ Greenwich Mean Time is named for the Royal Observatory at Greenwich, England. A discussion of Greenwich Mean Time is available at www.greenwichmeantime.com, as are conversions to local time zones.

¹⁸ Internally, the database software distinguishes between dates on which a variable is not defined (such as the Federal Reserve’s M2 monetary aggregate prior to 1959) and dates on which the series is defined (that is, was visible to observers monitoring the series at that date) but for which values are missing because, for example, certain printed publications cannot be located.

¹⁹ The database design is due to George Essig, senior web developer, Federal Reserve Bank of St. Louis.

Table 1**Revision History for 2004:Q1 GDP, April 29 through December 29, 2004**

Value	Variable name	Measurement interval	Validity interval	Transaction interval
11447.8	GDP	{1Jan2004, 31Mar2004}	{29Apr2004, 26May2004}	{29Apr2004, 31Dec9999}
11459.6	GDP	{1Jan2004, 31Mar2004}	{27May2004, 24Jun2004}	{27May2004, 31Dec9999}
11451.2	GDP	{1Jan2004, 31Mar2004}	{25Jun2004, 29Jul2004}	{25Jun2004, 31Dec9999}
11472.6	GDP	{1Jan2004, 31Mar2004}	{30Jul2004, 28Jul2005}	{30Jul2004, 31Dec9999}
11457.1	GDP	{1Jan2004, 31Mar2004}	{29Jul2005, 31Dec9999}	{29Jul2005, 31Dec9999}

Transaction intervals show dates on which Federal Reserve Bank of St. Louis staff entered or changed data values. The first date is the date on which the datum was added to the database. The ending date of the interval is infinity (open-ended) and will remain so indefinitely unless the datum (in the first column) is erroneous. Erroneous values in the database never are changed or removed—doing so would destroy the database’s historical integrity. Once a datum has been made visible to web site customers, integrity of the database requires that the row is never modified or deleted. Instead, erroneous values are corrected by adding an additional row to the database for the same measurement and validity intervals. When a new row is added to correct an error, the end date of the erroneous row’s validity interval will be set to the day on which the new row is added, and the start date of the new row’s transaction interval will be set to the same date.²⁰ A customer selecting a date interval that includes the correction date will receive both the original erroneous datum and the corrected datum, plus a message warning that the observation for that date was corrected. The customer is responsible for checking his empirical results using both values.

The initial version of ALFRED will not include transaction intervals. This omission matters not at all so long as data never are changed after being

made visible to researchers via the Internet. Because initial data will be machine-loaded from archival files, data entry errors and corrections are unlikely. Programming ALFRED using the three-interval architecture is significantly more difficult than with a two-interval (measurement, validity) design and would significantly lengthen ALFRED’s development.

The ALFRED project will make it unnecessary to archive datasets for studies based on data obtained from FRED so long as the author retains the RDS signatures for the dataset.²¹ Yet, what of the careless or forgetful researcher who does not retain the RDS? For them, the current design includes automatic archiving and retrieval of RDS data if the researcher signs up for a user account. After so doing, they will be offered the opportunity to save, on FRED and ALFRED, the RDS strings for every series they download. Putative replicators need only ask the original researcher to retrieve the RDS strings and make them available.

The ALFRED system, when completed, promises to unify the concepts of real-time data and replication in economic research.

²⁰ Once created, the integrity of the database requires that every row be retained in the database; otherwise, entering the same data signature string on different dates would retrieve different data, which is unacceptable. Including a transaction interval is an essential design element in a database system that guarantees time-invariance of retrieved data when the data are subject both to revision by the publisher and to possible human data-entry error.

²¹ An additional, related part of the FRASER/ALFRED project, nearing completion, is a catalog of available federal, state, and local data series. In its intent and structure, the catalog resembles the “Statistical Knowledge Network” discussed by Hert, Denn, and Haas (2004) except that the catalog does not attempt to explain or instruct in the ways that the data might be used to conduct economic analyses, as these authors suggest their metadata might be able to do. Also, to the extent that descriptive metadata is stored as XML tags, the design of Hert, Denn, and Haas is compatible with the VDC/DDI initiative to cross-index data from various servers on the Internet. Further, since the St. Louis economic data catalog will index data from all government agencies, it partially circumvents the barrier to cross-government IT collaboration discussed by Mullen (2003), although the *GPO Access* project (U.S. Government Printing Office, 2001) has been charged by the Congress to promote electronic dissemination of data and documents.

THE VDC PROJECT

The VDC project of Harvard University, similar to ALFRED, has as its goal increasing the replicability of research.²² Unlike ALFRED, however, the VDC project itself does not include data collection. Rather, the heart of the VDC project is to provide a low-cost, integrated suite of software that will allow other researchers a forum for archiving and sharing data. More precisely, the VDC project furnishes “an (OSS) [open-source software] digital library system ‘in a box’ for numeric data” that “provides a complete system for the management, dissemination, exchange, and citation of virtual collections of quantitative data.”²³

An essential component of the VDC project’s architecture are the tools to encourage researchers—including individuals, professional journals, and research institutions—to use a single set of formatting and labeling standards when they place datasets on the Internet. In turn, a loosely coupled web of VDC servers will locate, index, and catalog the datasets, making them available to other researchers. The VDC’s proposed formatting and labeling standards are those of the University of Michigan’s Data Documentation Initiative (DDI) project, an accepted standard in the document and knowledge management arena.²⁴ The formatting consists solely of inserting plain text XML tags within text data files, easily done within many programs or a simple text editor.

If successful, the VDC project promises the type of network effects, well-known to economists, that accompany (and drive) the adoption of standards. Because each VDC node maintains a catalog of materials held on other VDC nodes, as the VDC network expands, additional researchers will find it increasingly attractive to join, so as to make their work visible to the growing community.

²² Altman et al. (2001, p. 464).

²³ See <http://thedata.org/>. The VDC and St. Louis projects share the same core open-source components: Linux, Apache, and PostgreSQL. The St. Louis middleware is coded as server-side PHP scripts, similar in spirit if not code to the Java servlets used in VDC.

²⁴ On DDI, see Blank and Rasmussen (2004) and www/icpsr.umich.edu/DDI/.

A COMPARISON OF PROJECTS: ALFRED AND VDC

Both the VDC and ALFRED projects provide tools that promise to improve the scientific quality of empirical economic research. But, their philosophies and architectures differ. Altman et al., for example, writes that “The basic object managed in the [VDC] system is the *study*” [emphasis added]. In the St. Louis ALFRED project, the basic objects managed are a published study’s set of signature strings, which permit repeated extraction of the same dataset from an underlying, encompassing database.

Because the VDC project focuses on preserving specific datasets from specific studies, it is well-suited for archiving both experimental and non-experimental data. The ALFRED project’s focus on archiving vintages of non-experimental data makes it better suited to macroeconomic research, both real-time and replication studies. In *replication* studies, the issue is determining which data were used in a study and whether the calculations were performed as described; in *real-time* data studies, the issue is determining the robustness of the study’s findings to data revisions. At least for aggregate macroeconomic data, an archival system that can do two things—provide a later investigator with the previous researcher’s original data *as well as* provide earlier and later published values of the same variables—has the promise of combining a “simple” replication study with a real-time, data-based robustness study. In Pesaran and Timmermann’s notation, the archival system must be able to produce, on demand, both the universe of all observations on the variables of interest, χ , and all possible time-indexed “real time” subsets, N_{xt} . Although careful use of XML tags in a VDC/DDI system might permit support for such real-time econometrics, it likely would require attaching XML tags to each data point in each study.

CONCLUSION

Data archiving, data sharing, and replication are hallmarks of science, necessary to explore the correctness of published results. Real-time data

studies are important to address the robustness of published results. These two lines of inquiry are linked by means of their recognition that empirical economic research is inherently time-indexed. Although quite different, the VDC and ALFRED projects promise to assist and improve the quality of empirical economic research by reducing the cost of both lines of inquiry.

REFERENCES

- Altman, Micah L.; Diggory, Andreev M.; King, G.; Sone, A.; Verba, S.; Kiskis, Daniel L. and Krot, M. "A Digital Library for the Dissemination and Replication of Quantitative Social Science Research: The Virtual Data Center." *Social Science Computer Review*, 2001, 19, pp. 458-70.
- Anderson, Richard G.; Greene, William H.; McCullough, Bruce M. and Vinod, H.K. "The Role of Data & Program Code Archives in the Future of Economic Research." Working Paper 2005-14, Federal Reserve Bank of St. Louis, 2005.
- Anderson, Richard G. and Dewald, William G. "Replication and Scientific Standards in Applied Economics a Decade After the Journal of Money, Credit and Banking Project." Federal Reserve Bank of St. Louis *Review*, November/December 1994, 76(6), pp. 79-83.
- Bailar, John C. "The Role of Data Access in Scientific Replication." Presented at the October 16-17, 2003 workshop "Confidential Data Access for Research Purposes," held by the Panel on Confidential Data Access for Research Purposes, Committee on National Statistics, National Research Council, 2003; www7.nationalacademies.org/cnstat/John_Bailar.pdf.
- Bernanke, Ben S. and Boivin, Jean. "Monetary Policy in a Data-Rich Environment." *Journal of Monetary Economics*, 2003, 50, pp. 525-46.
- Blank, Grant and Rasmussen, Karsten Boye. "The Data Documentation Initiative." *Social Science Computer Review*, Fall 2004, 22(3), pp. 307-18.
- Bornstein, Robert F. "Publication Politics, Experimenters Bias and the Replication Process in Social Science Research," in James Neuliep, ed., *Replication Research in the Social Sciences*. Thousand Oaks, CA: Sage Publications, 1991, pp. 71-84.
- Boruch, Robert F. and Cordray, David S. "Professional Codes and Guidelines in Data Sharing," in Stephen E. Fienberg, Margaret E. Martin, and Miron L. Starf, eds. *Sharing Research Data*. Washington DC: National Academy Press, 1985; www.nap.edu/openbook/030903499X/html/index.html.
- Christoffersen, Peter; Ghysels, Eric and Swanson, Norman R. "Let's Get 'Real' About Using Economic Data." *Journal of Empirical Finance*, 2002, 9, pp. 343-60.
- Clark, Todd E. and Kozicki, Sharon. "Estimating Equilibrium Real Interest Rates in Real Time." Working Paper 04-08, Federal Reserve Bank of Kansas City, September 2004.
- Crawford, Vincent P. and Sobel, Joel. "Strategic Information Transmission." *Econometrica*, November 1982, 50(6).
- Croushore, Dean and Stark, Tom. "A Real-Time Data Set for Macroeconomists." *Journal of Econometrics*, 2001, 105, pp. 111-30.
- Dewald, William G.; Thursby, Jerry G. and Anderson, Richard G. "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project." *American Economic Review*, September 1986, 76(4), pp. 587-603.
- Faust, Jon; Rogers, John H. and Wright, Jonathan H.. "Exchange Rate Forecasting: The Errors We've Really Made." *Journal of International Economics*, May 2003, 60(1), pp. 35-59.
- Federal Reserve Bank of St. Louis. "As It Happened: Economic Data and Publications as Snapshots in Time," presentation by Robert Rasche, Katrina Stierholz, Robert Suriano, and Julie Knoll, at the Fall Federal Depository Librarians Conference, October 19, 2004, Washington, DC; fraser.stlouisfed.org/fdlp_final.pdf.

Anderson

- Feigenbaum, S. and Levy, D. "The Market for (Ir)Reproducible Econometrics" and "Response to the Commentaries." *Social Epistemology*, 1993, 7(3), pp. 215-232 and pp. 286-92.
- Fienberg, Stephen E.; Martin, Margaret E. and Starf, Miron L., eds. *Sharing Research Data*. Washington, DC: National Academy Press, 1985; www.nap.edu/openbook/030903499X/html/index.html.
- Goldberg, Lawrence G. and Saunders, Anthony. "The Growth of Organizational Forms of Foreign Banks in the U.S." *Journal of Money, Credit, and Banking*, August 1981, pp. 365-74.
- Hert, Carol A.; Denn, Sheila and Haas, Stephanie W. "The Role of Metadata in the Statistical Knowledge Network." *Social Science Computer Review*, Spring 2004, 22(1), pp. 92-99.
- Kishor, N. Kundan and Koenig, Evan F. "VAR Estimation and Forecasting When Data Are Subject to Revision." Working Paper 2005-01, Federal Reserve Bank of Dallas, February 2005.
- Koenig, Evan F.; Dolmas, Sheila and Piger, Jeremy. "The Use and Abuse of Real-Time Data in Economic Forecasting." *Review of Economics and Statistics*, 2003, 85, pp. 618-28.
- Mullen, Patrick R. "The Need for Government-Wide Information Capacity." *Social Science Computer Review*, Winter 2003, 21(4), pp. 456-63.
- National Research Council. *Access to Research Data in the 21st Century: An Opening Dialogue Among Interested Parties*. Report of a workshop on the Shelby Amendment held by the Science, Technology and Law Panel of the National Research Council, March 12, 2001. Washington, DC: National Academy Press, 2002.
- Neely, Christopher J.; Roy, Amlan and Whiteman, Charles H. "Risk Aversion versus Intertemporal Substitution: A Case Study of Identification Failure in the Intertemporal Consumption Capital Asset Pricing Model." *Journal of Business and Economic Statistics*, October 2001, 19(4), pp. 395-403.
- Orphanides, Athanasios. "Monetary Policy Rules Based on Real-Time Data." *American Economic Review*, September 2001, 91(4), pp. 964-85.
- Orphanides, Athanasios and van Norden, Simon. "The Unreliability of Output-Gap Estimates in Real Time." *Review of Economics and Statistics*, November 2002, pp. 569-83.
- Orphanides, Athanasios and van Norden, Simon. "The Reliability of Inflation Forecasts Based on Output Gap Estimates in Real Time." Working Paper 2003s-01, CIRANO, HEC, Montreal, 2003.
- Pesaran, Hashem and Timmermann, Allan. "Real-Time Econometrics." *Econometric Theory*, 2005, 21(1), pp. 212-31.
- Snodgrass, Richard T. and Jensen, Christian S. *Developing Time-Oriented Database Applications in SQL*. San Francisco CA: Morgan Kaufman, 1999.
- Stark, Thomas and Croushore, Dean. "Forecasting with a Real-Time Dataset for Macroeconomists." *Journal of Macroeconomics*, 2002, 24, pp. 507-31.
- Svensson, Lars E.O. and Woodford, Michael. "Indicator Variables for Optimal Policy." *Journal of Monetary Economics*, 2004, 50, pp. 691-720.
- Svensson, Lars E.O. and Woodford, Michael. "Indicator Variables for Optimal Policy Under Asymmetric Information." *Journal of Economic Dynamics and Control*, 2003, 28, pp. 661-90.
- U.S. Government Printing Office. *Biennial Report to Congress on the Status of GPO Access*. Washington, DC: U.S. GPO, 2001; www.gpoaccess.gov/biennial/index.html.
- U.S. Government Printing Office. *Report on the Meeting of Experts on Digital Preservation*. Washington, DC: U.S. GPO, March 12, 2004a; www.gpoaccess.gov/about/reports/preservation.pdf.
- U.S. Government Printing Office. *Concept of Operations for the Future Digital System*. Washington, DC: U.S. GPO, October 1, 2004; www.gpo.gov/news/2004/ConOps_1004.pdf.

APPENDIX

Recommendations of the Committee on National Statistics of the National Research Council and the National Academy of Sciences

The Committee on National Statistics' final report (Fienberg, Martin, and Starf, 1985) offered 16 specific recommendations regarding sharing research data. These recommendations, little noticed during the past 20 years, are as relevant today as then. Taken together, they form a foundation, or body of knowledge, for best-practice in empirical scientific research. Their recommendations are reproduced here because, although they sound scientific and sensible, most have been ignored in economic science.

For all researchers:

1. Sharing data should be a regular practice.

For initial investigators:

2. Investigators should share their data by the time of publication of initial major results of analyses of the data except in compelling circumstances.
3. Data relevant to public policy should be shared as quickly and widely as possible.
4. Plans for data sharing should be an integral part of a research plan whenever data sharing is feasible.
5. Investigators should keep data available for a reasonable period after publication of results from analyses of the data.

For subsequent analysts:

6. Subsequent analysts who request data from others should bear the associated incremental costs.
7. Subsequent analysts should endeavor to keep the burdens of data sharing on initial investigators to a minimum and explicitly acknowledge the contribution of the initial investigators.

For institutions that fund research:

8. Funding organizations should encourage data sharing by careful consideration and review of plans to do so in applications for research funds.
9. Organizations funding large-scale, general-purpose data sets should be alert to the need for data archives and consider encouraging such archives where a significant need is not now being met.

For editors of scientific journals:

10. Journal editors should require authors to provide access to data during the peer review process.
11. Journals should give more emphasis to reports of secondary analyses and to replications.
12. Journals should require full credit and appropriate citations to original data collections in reports based on secondary analyses.
13. Journals should strongly encourage authors to make detailed data accessible to other researchers.

For other institutions:

14. Opportunities to provide training on data-sharing principles and practices should be pursued and expanded.
15. A comprehensive reference service for computer-readable social science data should be developed.
16. Institutions and organizations through which scientists are rewarded should recognize the contributions of appropriate data-sharing practices.

