As It Happened:

Economic Data and Publications as Snapshots in Time

Federal Reserve Bank of St. Louis

Presented at the Fall Federal Depository Library Conference
October 19, 2004

Speakers:
Robert Rasche
Katrina Stierholz
Robert Suriano
Julie Knoll

Good morning.

As the introduction indicated, I am not a librarian. By training I am an economist; most of my worklife was spend in academia; for almost six years I have been the research director at the Federal Reserve Bank of St. Louis. My perspective on government documents is that of a long-time user: I can probably match anyone in attendance in the number of hours spent sitting on the floor of the stacks in government documents collections searching for obscure data.

Our presentation this morning is something of a tag team match. I will present some background on which government documents really interest the economic research community and why. Katrina Stierholz who manages the research library at the St. Louis Fed will address the question of why in the world a small Federal Reserve Bank research library got involved in a project to digitize and electronically disseminate selected government documents – in particular how we convinced anyone that this is an appropriate part of our mission. Rob Suriano, who supervises our FRASER project will talk about how we organize the production of the images, including scanning, cleaning, conversion to .pdf and OCR. Finally, Julie Knoll will bat cleanup and talk about the delivery systems that we use or are developing.

Economic researchers are interested in arcane government documents for two reasons, both related to the revision process through which many economic data evolve. First, in contrast to "hard sciences" such as physics and chemistry, standards of documentation in empirical research in economics are pretty slipshod. Historically it was common practice to send a research assistant or undergrad off to the library to collect the desired data series. Current practice usually involves extracting the desired information from any number of electronic databases, including the FRED collection of data that is provided on the Internet on our website. The sad joke in the economics profession is that it is virtually impossible to reproduce any empirical result that appears in print – peer-reviewed or otherwise. This was a long-standing bone of contention by my predecessor, a former journal editor, and one of my current colleagues, who wrote in the *American Economic Review* on the issue almost two decades ago.[1] Little has changed in the interim.

Second, modern theoretical models of economic activity, particularly models of the macroeconomy and the impact of fiscal and monetary policies, emphasize the role of expectational forces. Expectations are formed on the information known at the point-in-time, not on the revised information that becomes available after the fact. Hence to implement this type of model correctly it is necessary to have access to data points that are indexed in two dimensions: the point-in-time at which the data are published and the period of time to which the observation applies. As an example, consider the data that are published monthly on payroll employment. On October 8, 2004, an initial observation for September employment was released by the Bureau of Labor Statistics. It showed an increase of 96,000 employed persons – generally less than prognosticators anticipated. On the first Friday each of November and December a revised value will be published for this observation. Such revisions can be substantial. However, as we sit here in October, the only information we have available to utilize in forming expectations (i.e. making forecasts) about the future path of the economy is the initial estimate of

[1] Richard G. Anderson, William G. Dewald and Jerry Thursby, "Replication and Scientific Standards in Empirical Economics: Evidence from the JMCB Project," *American Economic Review*, September 1986.

September's employment, not the information that will be available to us about September come December or later.

For us, creating "digital master files" of historic data publications is what economists call "intermediate product" – a means to an end. What we are really interested in constructing and providing are "derivative files"[2] We are creating two different types of derivative files. The first are just .pdf versions of various scanned pages. We post these to our FRASER web site <fraser.stlouisfed.org>. The .pdf format is useful because it is easily accessible throughout the international economics research community. Searchable .pdf is a valuable asset to this community. We typically perform OCR on the scanned pages. OCR is not a 100 percent accurate process and since we do not have the resources to verify every number or character on every page that we post to FRASER one policy issue that we face is under what conditions we will allow search access to the pages. We will discuss our thoughts on this issue later in this discussion.

A second desirable aspect of the .pdf format is that we can organize materials in along different dimensions. For example, we have each monthly issue of *Economic Indicators* posted as a separate file on FRASER. The researcher who selects the file for a particular issue sees the available data at a point in time on many variables. In addition we have constructed files of individual tables (for example Federal Budget Receipts and Expenditures or Employment in Business and Government.) that appear regularly in *Economic Indicators*. These tables are posted in files grouped in five-year segments to keep the files to a reasonable size. Hence a researcher interested on how data on one

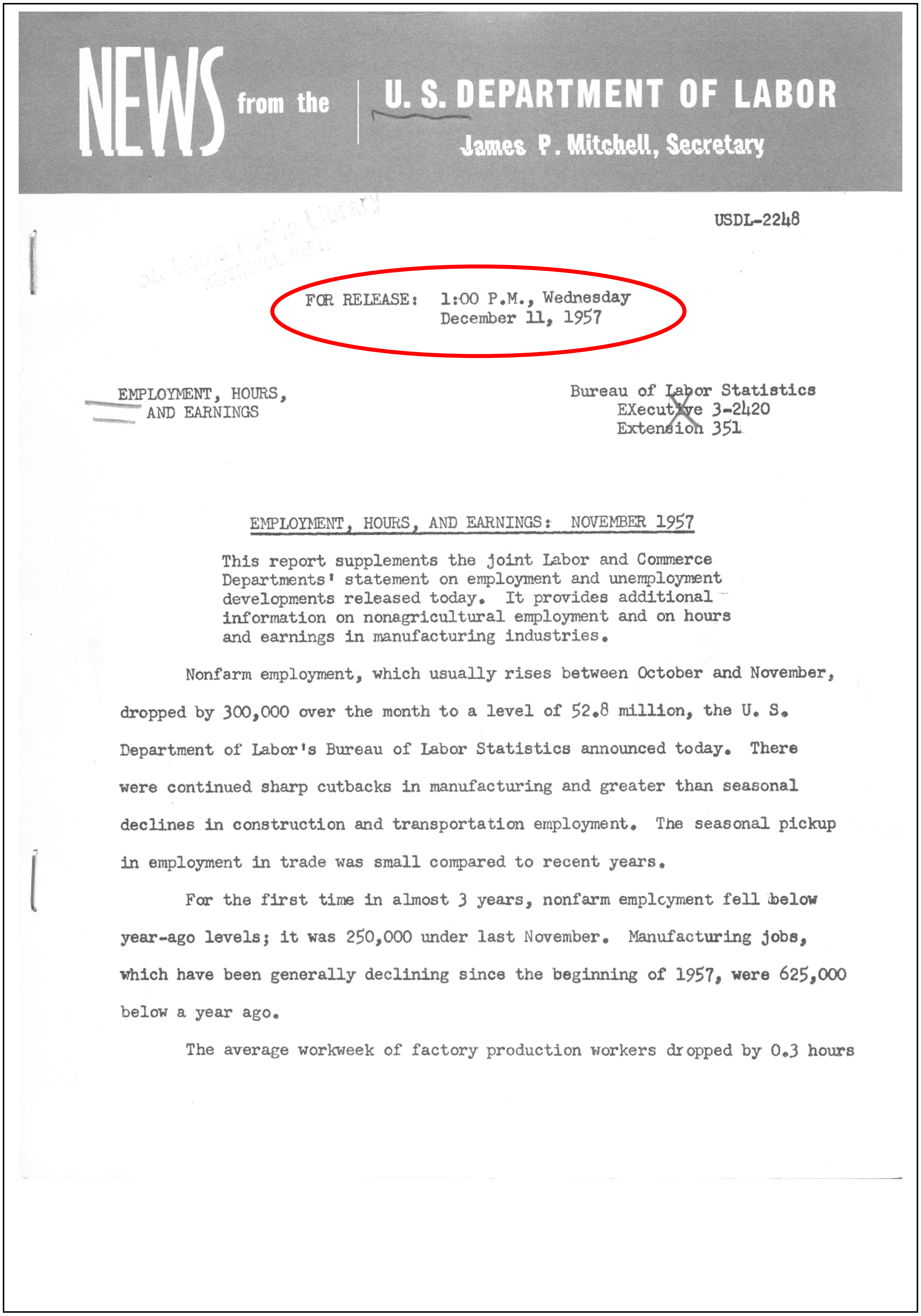
[2] *Report of the Meeting of Experts on Digital Preservation*, Washington DC: U.S. Government Printing Office, March 12, 2004. <www.gpoaccess.gov/about/reports/preservation.html>, p. 3

particular variable evolved over time can download these files and proceed through the tables chronologically.

The second type of derivative files will appear in a service that we will launch during 2005 – AlFred, for ArchievaL FRED. This site will allow a researcher to specify a set of data series and a range of dates, say 1-Jun-1974 through 30-Jun-1974. This inquiry will create either a text or an Excel file that contains all the observations on the specified series as they were available during the indicated interval. Since August 1999 we have created an archive of our FRED electronic database of economic data at the end of each month. We have extended these histories backward using information in hard copy data publications that the Bank prepared since 1966 and from other published sources, generally those of the originating agency. Unfortunately, we have found that in some cases when comprehensive revisions to data are released, the entire history has neither been published nor preserved by the originating agency. To the extent we have been able to locate information in secondary sources we have been able to fill in many of the gaps. Regrettably, it appears to be impossible to reconstruct parts of some histories.

We have assembled complete monthly histories, to the extent that the history has been preserved, for 24 series, generally beginning with the data that became available in January 1966. In one case, Industrial Production, we have complete histories beginning with May 1927. We are in the process of constructing histories for about 40 more series.

It is desirable to know the dates upon which various series were first made available to the public. Generally, since about 1994, these statistical releases are available on Web sites of the originating agency. For earlier information, in some cases we have been able to find the original news releases for the statistics that contain the date of release (see example.) If anyone has such documents squirreled away in cold storage, we would be happy to borrow them and tabulate the publication information. In other cases we rely on historic statistical publications of the Federal Reserve Bank of St. Louis. Starting in 1966 these publications these publications contained computer printouts with the most recent information on various data series. At the bottom of each computer printout was a date of preparation. Where we have the information to cross-check against the news releases we find that the printouts were prepared either on the date the data were released or the next business day. A third source of dating information is published "scheduled release dates" from the Department of Commerce and the Department of Labor. Generally these schedules are available starting around 1980. Finally, lacking any other source of date information, we use the last day of the month in which the updated data were released.



NEWS from the U. S. DEPARTMENT OF LABOR

James P. Mitchell, Secretary

USDL-2248

FOR RELEASE: 1:00 P.M., Wednesday
December 11, 1957

EMPLOYMENT, HOURS,
AND EARNINGS

Bureau of Labor Statistics
EXecutive 3-2420
Extension 351

EMPLOYMENT, HOURS, AND EARNINGS: NOVEMBER 1957

This report supplements the joint Labor and Commerce Departments' statement on employment and unemployment developments released today. It provides additional information on nonagricultural employment and on hours and earnings in manufacturing industries.

Nonfarm employment, which usually rises between October and November, dropped by 300,000 over the month to a level of 52.8 million, the U. S. Department of Labor's Bureau of Labor Statistics announced today. There were continued sharp cutbacks in manufacturing and greater than seasonal declines in construction and transportation employment. The seasonal pickup in employment in trade was small compared to recent years.

For the first time in almost 3 years, nonfarm emplcyment fell below year-ago levels; it was 250,000 under last November. Manufacturing jobs, which have been generally declining since the beginning of 1957, were 625,000 below a year ago.

The average workweek of factory production workers dropped by 0.3 hours

Series Name	First column	Last column	First observation
Civilian Employment	11-Jan-66	24-Sep-04	Jan-48
Exports of Goods and Services Constant	Jan-66	24-Sep-04	Mar-47
Exports of Goods and Services Current	Jan-66	1-May-04	Mar-46
Government Purchases of Goods and Services, Constant	Jan-66	1-Mar-04	Mar-47
Government Purchases of Goods and Services, Current	Jan-66	1-Mar-04	Mar-46
Gross Domestic Product, Constant Prices	Jan-66	1-Mar-04	Mar-47
Gross Domestic Product, Current Prices	Jan-66	1-Mar-04	Mar-46
Gross Private Domestic Investment, Constant	Jan-66	1-May-04	Mar-47
Gross Private Domestic Investment, Current	Jan-66	1-May-04	Mar-46
Imports of Goods and Services, Constant	Jan-66	1-Jun-04	Mar-47
Imports of Goods and Services, Current	Jan-66	1-Jun-04	Mar-46
Industrial Production	May-27	1-Jan-04	Jan-19
Output per Man Hour, Business Sector	Jun-68	1-Jun-04	Mar-47
Output per Man Hour, Nonfarm Business Sector	Jun-68	1-Jun-04	Mar-47
Payroll Employment	17-Jan-66	24-Sep-04	Jan-39
PCE, Constant	Jan-66	1-Aug-99	Mar-47
PCE, Current	Jan-66	1-Aug-99	Mar-46
Monthly Personal Income, Nominal	19-Jan-66	1-May-04	Jan-46
Monthly Personal Income, Real	Nov-76	Sep-04	Jan-48
Monthly Disposable Personal Income, Constant	Jan-79	Sep-04	Jan-48
Monthly Disposable Personal Income	Jan-79	Sep-04	Jan-48
Retail Sales, Nominal	17-Jan-66	1-Aug-99	Jan-47
Unit Labor Costs, Private Business Sector	Jun-68	2-Jul-04	Mar-47
Real Manufacturing and Trade Sales	Nov-76	Sep-04	Jan-48

The FRASER project, along with ALFRED and the union catalog, are a part of our mission, because providing economic data to the public has historically been an important mission for the St. Louis Federal Reserve Bank of St. Louis.

First of all, I'd like to explain a little bit about the Federal Reserve System and where the St. Louis Fed fits into the organization. The Federal Reserve System is divided into two parts: the Federal Reserve Banks and the Board of Governors. The Board of Governors is part of the federal government. The 12 Federal Reserve Banks are not. The Federal Reserve Banks operate as independently chartered banking institutions. I think the world sees us as a single entity, but we see ourselves as 13 very distinct institutions.

There is no "head Federal Reserve Bank". The BOG oversees each Federal Reserve Bank, but we operate very independently. Each bank has, over time, assumed a specialization. For instance, the Federal Reserve Bank of New York is in charge of open market operations and keeps an eye on international banking agreements. The Federal Reserve Bank in San Francisco focuses on the Asia-Pacific region. The Kansas City Fed specializes in issues in rural America. And the St. Louis Fed specializes in economic data.

One thing every Fed does have is economists. The economists' research various economic questions and offer advice to the president of the Federal Reserve Bank regarding the state of the economy and the federal funds rate set by the FOMC (aka, the "interest rate"). So, the Board of Governors and every Federal Reserve Bank has a research library to support these economists and their work. I work in the St. Louis Fed's research library.

When I joined the St. Louis Fed just a little over 2 years ago, Bob began this search for the economic data that would tell him exactly what economists knew at a particular point in time. A kind of “what did they know and when did they know it” question for economists and economic policy. We began by searching for press releases put out when various economic indicators were released—CPI, PPI, Retail Sales—pretty much anything that offered release dates to add to the ALFRED database. Bob saw the value of this information, and decided to make it available in two ways—as the publication (a scanned document via FRASER) and as a database with data points and release dates (ALFRED). This would build on our strength as the economic and financial data source. As Bob mentioned, some of the information was in our own publications, but much of what we’ve gotten has come from depository libraries and the help of depository librarians. I will talk a little later about some of our hurdles in getting this information, as well as what I see as issues for libraries to deal with when it comes to retaining information.

I have a personal experience with economic data that really brought home for me the problems of that 20/20 hindsight. As a product of getting some material together for Bob, I retrieved the National Summary of Business Conditions published in the Federal Reserve Bulletin from about 1927 to the early 1940’s. These were commentary—a page or two, describing the state of the economy. Well, of course I started reading them. Because the Depression was such a profound influence on my parents, I read especially carefully the months following the crash in 1929.

[Industrial production declined further in October, and ther was also a decrease in factory employment. As compared with a year ago, industrial activity

> continued to be at a higher level, and distribution of commodities to the consumer was sustained. Bank credit outstanding increased rapidly in the latter part of October, when security prices declined abruptly and there was a large liquidation of brokers' loans by nonbanking lenders. In the first three weeks of November further liquidation of brokers' loans was reflected in a reduction of security loans of member banks. Money rates declined throughout the period.
>
> First paragraph, National Summary of Business Conditions published in the *Federal Reserve Bulletin,* Dec 1929]

I had expected to read something earth shattering—something close to the sky is falling. I kept thinking—can't they see what's coming? Don't they know it's the Great Depression? But, no, they didn't. It is with 20/20 hindsight that we realized what was coming—how bad it would be and how long it would last.

A little more about the St. Louis Fed

Back to St. Louis and our specialization in economic data. Every Federal Reserve Bank gathers data. What makes the St. Louis Fed unique is that we also publish data. We are, and have been for almost 40 years, a premier provider of economic data. And the reason for publishing data comes from an economic debate, a debate that St. Louis played a large role in arguing. The data was essentially published to prove a point, to back up an argument. The case that was being made was in support of an economic argument between Monetarists and Keynesians. An argument over what controlled inflation. Economic data would be an important tool for economists in this debate.

Many economists at the St. Louis Fed were Monetarists. Monetarists believed that the money supply is an important measure/indicator of the economy and that the supply of money directly influenced inflation. The St. Louis Fed was famous for its advocacy of Monetarism, which was counter to what most of the economists at the BOG and other Feds thought.

At the same time that St. Louis became a home for Monetarists, it also began to change in other ways. The head of Researcch at the time was Homer Jones (I work in the Homer Jones library). Mr Jones changed the Review (our bank's flagship publication) during this time. Until then, articles were unsigned and had no footnotes. During Homer Jones' time, articles began to list authors and have footnotes. The Review went from being a local bank publication with a focus on regional economics to an academic-style publication with a focus on national economics. As the head of Research, Jones required the authors of articles in the Review to back up their statements with empirical data.

So, the focus on monetarism, combined with these academic-level expectations of research and writing, and the emphasis on empirical work, brought about this focus on economic data. To support their arguments (and to provide all economists of any policy persuasion the same ammunition), the St. Louis Fed began publishing data series for economic researchers.

Data publications produced (these became the basis for FRED)

Triangles of US Economic Data (monthly, Jan 1966-May 1967)

Triangles of US Economic Data: Quarterly Supplement (Jan 1966-May 1967)

USFD (weekly, January 1966-)

Monetary Trends (monthly, June 1967-)

National Economic Trends (monthly, August 1967 -)

These publications were freely distributed, and added to the empirical information available as to cause and effect for Fed policies and the economy

So, as a result of the St. Louis Fed's focus on backing up their answers to questions with data, the St. Louis Fed collected, organized, and produced data. The intent was to provide empirical evidence for researchers to study, and then use as support for their ideas. The fundamental goal was to provide high quality economic and financial data.

For 15 years, the data was only produced in paper. Thirteen years ago FRED was born. FRED has been 'live' or 'alive' now for 13 years.

FRED

We have very little information on FRED as a baby —a reflection of how hard it is to see the need to archive information at the moment the information is created (another 20/20 hindsight). I have an email from someone describing it when it was a BBS (the email is from 1993, FRED went live late in 1991). See powerpoint slide announcing electronic bulletin board (baud, etc.). And I have a screen shot of FRED from 1996, after it had gone live on the web. Both of those images are from the wayback machine, not from our files (oops). When we went online, more data was added. Even the print publications grew, but the electronic version grew

FRED/FRASER

So, that's the history of FRED and how our focus on data all started. Rob will discuss our current FRED, FRASER, and the projects that we're working on.

LIBRARY ISSUES

Before I turn it over to Rob, I want to spend a few minutes on some library issues. While I was looking for the material to fulfill our need to have release dates for the data, I ran into an interesting problem. I had a very hard time finding the release date for data. Government agencies put out press releases, with the initial data. Then the revised data would come out, often as part of the next month's press release. When the third and final release was done, a real document was released (e.g., the monthly CPI publication). Virtually every library threw out the press releases, as these were considered superseded by "better" and newer data. The problem with that is that you lose the sense of what was known at the time. In many cases, I called documents librarians all over the country, searched online catalogs for so many libraries (thank God for consortium catalogs), and even called the issuing agency and the Library of Congress. Often, it was a depository library that came through, when both LC and the agency could not. It is interesting to me that even with the FDLP's strict retention requirements, so few libraries kept the information. And, for that matter, agencies often did not keep the material either.

Besides losing the print press releases, or other "current news" that was subsequently revised, the electronic versions were also often lost. We were as bad about this as anyone. When data was received, a new version was written over the old version. Backups were maintained, but for the purposes of recreating the current data—not for keeping a historical record of the data. There have been several cases where I've called the issuing agency to see if they have electronic versions of their old data (how easy would that be for us?). Unfortunately, because disk space was scarce, files were overwritten. I have not yet gotten any significant old data in electronic form. We were just as bad, but are now reformed. In 1999, the St. Louis Fed began taking weekly snapshots of our data. In the electronic case in particular, the loss is permanent. While we would often find the paper press releases in libraries as a result of benign neglect, for electronic files there was only one copy, and they were all overwritten.

As librarians, it is so very hard to predict what users will need in the future. Libraries have such pressure to remove material that appears to have no value--and one could reasonably argue that these data releases had no value. But, in the end, there is great value to their initial judgments. Policy decisions are made with the data that is released at the moment, not with the perfect data that we see now. It is important to see the data using the same imperfect lens as economists had at the time it was released.

One other important point (and a great big thank you)—we have benefited so much from government documents libraries and librarians. Much of what Rob will tell you about with regards to FRASER has only been possible through the generosity of depository librarians. St. Louis Public let us take all of their unbound Economic Indicators, remove the staples, scan them, and then re-staple them and give them back. We have also gone through the Needs and Offers lists and requested items that are now being scanned. I am truly grateful, but not surprised, at the generosity and helpfulness of so many documents librarians who searched their shelves for hidden treasures, let us borrow and copy, and for those who sent their Needs and Offers list to the National list. We have made great use of that list. As my personal 2 cents—if you have a lovely, long run of an economic item, please feel free to give me a call or drop me an email. I will take them off your hands and refund your shipping costs.

So, in the end, why is this an appropriate part of our mission?

1. It serves to further economic knowledge, a core mission for the St. Louis Fed
2. We continue our work in giving economists and researchers easy access to a wide variety of data. We intend to continue to be a single, authoritative site for finding economic data.

With that, Rob will now come up and describe FRED and FRASER.

Economic Data at STLFRB

As Katrina and Bob mentioned, our goal at the Federal Reserve Bank of St. Louis is to become the premier sources of economic and financial data. I want to spend most of my time discussing our newest resource – FRASER – but I want to take a few moments to talk about our data publications and FRED, both of which are the mainstay of our resources. Then, after I discuss FRASER, I'll talk briefly about a few future projects that we hope will only enhance our reputation, as well as expand upon what we are already providing.

Data Publications

As a base for both economic theory (our Monetarist roots) and a source of data about both the economic conditions both regionally and nationally, the Bank has been a producer of reports, publications, and papers.

Posted on the web site are such publications as the Regional Economist and our nationally recognized Review. In addition there are statistics-based reports such as Monetary Trends, National Economic Trends, International Economic Trends, and U.S. Financial Data. Each of the latter publications graphically depicts the data that we make available in FRED and widely used sources in the financial, economic, and academic communities.

FRED

Our largest offering is FRED (or Federal Reserve Economic Data), which is an electronic database located on our web site. We currently offer over 3,000 economic time series on FRED, which include banking & financial, employment & population, monetary aggregates, interest rates, and US Trade data (among others) from the Board of Governors, the US Government, as well as commercial sources. Most series are available back to the mid-1960s to mid-1980s, with a few sereis – like the amount of currency in circulation – available back to before 1920.

The data on FRED is available for download to straight text, spreadsheet, or viewable online in graphical format. Each series we make available includes its source, when it was last updated, as well as any notes that may be pertinent to the user (such as change in the source or how the data was collected or aggregated).

In addition to just posting the data online, we also provide an email notification service that lets researchers know when a desired data series has been updated. This alleviates the need to continually check the site to see if a particular series has been released.

FRASER Screenshot

Our latest project – which went live July 1 of this year – is FRASER. (If you haven't had the chance to visit the site, this is what it looks like today. We even have a little survey up to let us gather feedback from users. We're actually demonstrating FRASER, as well as the other features of our site at our booth here at the conference, so please drop by.)

FRASER

FRASER is the Federal Reserve Archival System for Economic Research. It is an electronic archive of historic economic statistical publications.

Ultimately it will provide an essential link between those data series now available electronically (and particularly on FRED) with the historic past of those series currently only available in printed form. Such information as banking statistics from 1896 or consumer prices to the 1930s.

We are limiting the scope of publications we include on FRASER to government documents, and those in the public domain. We currently have no plans to include those with copyright restrictions. And, as we have let many of you who have asked, this a free site; there is no charge to access the publications on the site.

We currently have a modest number of publications available, focused at this time on national banking and economic data, but our long term goal will be to expand this to include other prominent national economic serials, such as Business Conditions Digest, and regional data, such as the City & County Data Book.

While we do have some data series on FRED that go back to the early part of last century, these are data that typically represent the final revised figures as presented by the originating organization. And, not all of the data on FRED goes back beyond the 1970s or 1960s. The underlying purpose of FRASER will be to make available data series as they were released – in their preliminary, revised, and final values. As you can imagine, this will ultimately extend the data on FRED to the past. And we hope someday to efficiently extract data from scanned images and provide seamless data series and make them fully useful to researchers.

I'd now like to take a few moments to talk about how we put FRASER together…

FRASER Workflow

This is a diagram illustrating the different steps we are using at the Federal Reserve Bank of St. Louis to create digital images from paper-based publications.

Each publication is scanned to produce Tagged Image File Format files. The file is cleaned of imperfections and saved as a multi-page TIFF file. The original image is archived while the “cleaned” image is run through an OCR routine to create underlying text and convert the file to PDF format for posting to the FRASER site.

We are also storing the original paper publications in the event that we have to go back and repeat the imaging. The original TIFF files are archived for the same purpose – in the event that we need to create a new PDF file.

Image Preparation

The imaging of a publication or document is dependent upon a number of factors, including the quality of the printing and the efficiency of the scanner. Also, depending on how the publication was utilized in its life on the shelf, there may be handwritten notations or underlining within the text, and some pages may have become smudged or even torn.

Our cleaning process – via the use of an editing software package, allows us to remove many of these non-published marks and imperfections.

Imaging imperfections include lines or spots due to dust or residue on the scanner lenses or rollers.

Misalignment is caused by pages not scanned perfectly straight. These consequently need to be reoriented to vertical.

The handwritten notations and library stamps that we run across are also removed from the scan as efficiently as possible. On occasion, this is not always possible, but our goal is to create an image that is as close to the original printed publication as possible.

The cleaning process is the most labor-intensive step. The software we are using – and I'll let Julie talk more about that aspect – is such that it allows for manual "erasing" of imperfections rather than automated removal. We've found that typically a person can clean around 100 pages or so an hour if the image is fairly clean of marks. The "dirtier" the image, the slower the rate of cleaning.

FRASER Production Status

As of this month – and in actuality last week – we had completed scanning of over 70,000 pages for 9 different economic publications. Almost half of our scanning has been completed in the past 3 months.

We have cleaned, OCR'd and posted some 38,000 pages and posted these to the FRASER site. Another 6,000 pages have been cleaned since June and are awaiting the completion of publication series before being moved to the web.

Right now we are working on the Business Conditions Digest series. It is a monthly publication with over 25 years of releases. We are looking to finish the 1960s in the coming weeks and post these to the site.

As a benchmark indication of how our process has proceeded, since June we have logged almost 600 hours in scanning and cleaning activities which has resulted in a cost per page of around 8 cents. This compares favorably to the cost to outsource this function, which we found to be from 25 cents to over a dollar per page back in 2003, depending on the number of pages scanned. And not all of the companies we received quotes from could scan at our specs.

Future Projects

In the long run, we are looking at a number of other offerings that will further enhance our position as a premier source for economic data.

Included, is the introduction of metadata to the files posted to FRASER that will enhance their usefulness. It would be great to be able to say this was already part of the FRASER process, but we've actually been waiting for the GPO to finalize its standards instead of going ahead with one of our own making.

As mentioned earlier, ALFRED, or Archival FRED, is an archival database of what we have posted to FRED over the years. Our practice has been to save the data on FRED each week onto DVD. By using the data saved to these discs, we will create a database that will allow researchers to see exactly what data was available on a given date, or date range, and allow them to analyze the effects of the revision process on economic models and lead to more effective forecasts and explanatory modeling. As an aside, and to those of you who might be interested, our principal programmer on this project has been greatly influenced by the text, Developing Time-Oriented Database Applications in SQL, by Richard T. Snodgrass which, unfortunately is out of print, but the author has made it available on his web site (http://www.cs.arizona.edu/people/rts/tdbbook.pdf). The information learned from this source has proved valuable during our development of ALFRED.

We are also working on the development of a Union Catalog of electronic resources related to economic information. The catalog will permit the user to search for economic reports, data releases, and information related to economics that are posted to the Internet around the world – including other central bank and U.S. agency web sites. We think this will provide a valuable referencing tool to economists and researchers.

I’ll turn this over now to Julie Knoll, who will tell you more about the technical aspects of FRASER related to the scanning, OCR and the web site.

Speaker: Julie Knoll

I'll be talking to you today about the hardware and software used for creating the documents and the website, some considerations to keep in mind, our back up procedures, and last how the data is actually input into the database.

Scanning and Preparation Tools

We started out using two different scanners. The first was a Xerox DocuCentre, which is a copier/scanner/printer. It was used to scan all of the Economic Indicators. We soon realized how time consuming this was and that we could not permanently take over the department's main copying machine. We purchased a Kodak i260 document scanner, and a second one soon after. This scanner has both a document sheet feeder and a flatbed screen. The scanner worked wonderfully at first, but we did start to have problems. Strange lines would appear throughout the documents and would not always disappear after cleaning the scanner and a part broke with no replacement that we could order. To say the least, we now have in our plans to purchase a Fujitsu scanner.

The Kodak scanners came with Kodak Capture Software that we use for scanning the documents. The software allows scanning at resolutions up to 2400 dpi, black-and-white, grayscale or color scanning, and outputs images to a variety of formats. It has a feature for customizable templates to save sets of configurations for different projects. It can also take an image of a book scanned from the flatbed and split the image in half, so that each page of the book becomes a separate image.

We scan all of our documents as 600dpi, black-and-white, multi-page TIF files. Scanning in grayscale or color is only done when specifically needed for a document. We find that multi-page files are easier to work with and to keep organized than are single-page files. Many software packages can split the document into single-pages files at any time.

ScanSoft PaperPort is then used to clean up the images. Black borders, handwritten comments, and other non-essential marks are erased from the pages. All pages are de-skewed and rotated to the correct orientation. This software works well for the most part. It is user-friendly and has pretty much all of the features we need. It has been known to use up a lot of the computers memory though.

OCR, or optical character recognition, is next. Up to now, ABBY FineReader was used for all documents. Other software was evaluated, and FineReader seemed to perform the best. Obviously there is no single software package that does equally well across different types of documents. FineReader was tested with Economic Indicators and did a reasonable job with nested table cells.

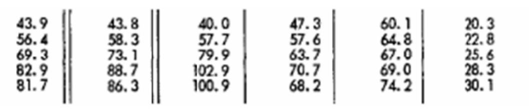
The OCR software exports the document to a PDF format containing the page images overlaid with clear text. This means that when you are looking at the PDF you are seeing the actual scanned image, but you are still able to select and search through the overlaid text. You can also use the software to export the text into a variety of other formats including Microsoft Word and Excel.

OCR Considerations

When performing OCR on a document there are multiple factors that must be considered. The quality of the original is the most important. The cleaner your original document is the more accurate the OCR will be. You should try to use the cleanest copies available, but when you're dealing with documents 100 years old, take whatever you can get.

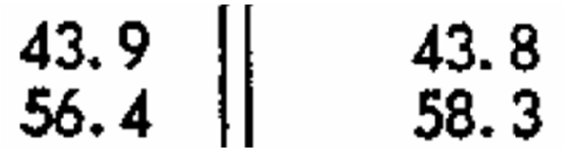
Scanning at high resolutions is best, but it will also create a large file size. Remember that the smaller the print, the higher resolution that will be needed. We scanned all of our documents at 600dpi, but during the PDF conversion stage, the images are reduced to 300dpi. Even still, this creates a rather large file for most documents. We decided to stay with this because the text is extremely small in some cases and we wanted users to be able to zoom in as much as needed and still have the text legible. These examples were taken from a finished PDF for Business Statistics. You can see here how small the text is when viewed at 100% compared to viewing it at 400%.

Business Statistics, 100%



43.9	43.8	40.0	47.3	60.1	20.3
56.4	58.3	57.7	57.6	64.8	22.8
69.3	73.1	79.9	63.7	67.0	25.6
82.9	88.7	102.9	70.7	69.0	28.3
81.7	86.3	100.9	68.2	74.2	30.1

Business Statistics, 400%



43.9	43.8
56.4	58.3

Your documents will not be 100% accurate. Even the best OCR software with the cleanest documents is only rated to be 99.9% accurate. This means that one out of every thousand words will be identified incorrectly.

As of yet, the OCR'd text for documents on the FRASER site has not been proofread or corrected. This is both a time-consuming and tedious process. Eventually, we do plan on correcting and extracting the data and placing in AlFRED, but for the time being, the main purpose of OCR for us is so that the final PDF will be full-text searchable.

One of our concerns is that a user might extract the text from the document without realizing that the data is incorrect. The way we prevent this is by implementing PDF security using Adobe Acrobat® on all PDF files. PDF security uses a password to restrict different features of the document. You are able to prevent users from opening the document, printing the document, and various levels of editing unless they have the password. We use the security to restrict copying and extracting of all text within the document. The text is still readable, and therefore still searchable by users and internet search engines, and it still accessible by screen reading software.

Website Software

On the FRASER, FRED, and eventually AlFRED sites we use a combination of Red Hat Linux, Apache web server, PostgreSQL database, and PHP programming language. These are all open-source projects. We have found the combination to be stable, secure, and fast.

PostgreSQL

We chose to use PostgreSQL for the database rather than the more popular MySQL because of the more advanced features that PostgreSQL offers such as transactions and foreign key relationships.

Transactions allow you to rollback a series of updates if an error occurs in the middle. If you are updating two related records and an error occurs during the second update, chances are you do not want to commit the first update. Transactions will only commit the updates if all updates are successful. Otherwise, it will rollback to the state of the database as it existed before the transaction began.

Foreign key relationships help with data consistency when you have two related tables. If a record in one table depends on another, a foreign keep will check that the proper record exists in the main table before any data is added to the subsequent table.

PostgreSQL also has many other features such as subqueries, views, and stored procedures that make application development much cleaner and simpler.

PDFLib+PDI

We also use an extension to PHP called PDFLib+PDI. PDFLib is a development tool allowing automatic PDF generation on the web server. The PDI piece, is available only with the commercial version of the software, and is a companion tool that allows you to process existing PDFs. We use the software to create the grouped page files on the site. It loops through the PDFs, extracts the correct page from each issue, and groups them together in one new file. These files are created beforehand and stored on the server. We also use the software to allow users to extract single pages out of any issue. These files are created on the fly, only when the user specifies a particular page.

Back-Up Procedures

Both FRED and FRASER use similar back up procedures. We use a combination of CVS, tape backups, contingency servers, and hard drive failover technologies.

CVS is a version control system that allows you to record the history of files. Every night any changes to documents and static HTML documents are recorded. This allows us to revert to a previous version of a file at anytime. In addition, the database for both sites is fully backed up each night by creating a dump file. A dump file is a text file that contains all the commands and data for rebuilding the database structure and restoring all of the data.

Each server is backed up to tape once a week. All documents and the database dump files are included in this. In case of server failure of data corruption, we will be able to restore the entire site from the tape backup.

We also have contingency servers at a remote location. Every night the main server is synchronized with the contingency server to create an exact replica. In the event of an emergency at our St. Louis location, we can switch service over to the contingency machine.

RAID (redundant array of independent disks) is a way of storing the same data in different places on multiple hard disks. If one drive were to fail, the same data could be read from another drive. A hot spare is a hard drive waiting to take over automatically and rebuild a failing RAID drive. Both servers have RAID 5 with hot spare. This means that we can have a hard drive failure and stay up and running with zero down time and no required human intervention.

For FRASER the original tiff files, the cleaned tiff files, and the final PDF files are all burned to CD or DVD.

Data Entry

The PDF files are not stored in the database. The files are stored regularly on the file system, and the path to the directory where the files are located is stored in the database. This and all other metadata is entered into the database by using simple web based administration forms. These forms are restricted to IP addresses within the bank and require a user name and password for access. The person doing the data entry must input all information about a document including the date, any keywords, table of contents, and the page number of the PDF for every page. These page numbers are used when creating the grouped page files and when extracting single pages from an issue. Each issue can be marked as private so another person can double check the information before it is publicly available.

This pretty much sums up the technical side of Fraser. I have included a slide that lists websites for all hardware and software mentioned. I will now open up the room if anyone has any questions.