Data Archiving at the U.S. Central Bank

by Linda F. Powell[1]

Abstract
As the central bank of the United States of America, the Federal Reserve System consumes vast quantities of economic, financial, and organization structural data. These data are used for making monetary policy, conducting banking supervision, performing economic research, and implementing consumer protection policies. The focus of this paper is on micro data archived at the Board of Governors of the Federal Reserve System. The paper discusses the types of data used by the central bank, how data are collected and edited, data documentation and meta data, and data purchased from commercial vendors. The paper discusses the challenges faced by archiving a diverse pool of data including communication and coordination, user access across various computer platforms, and meeting the diverse needs of a variety of end users. Finally, the paper discusses some of the solutions to the challenges faced and how technology is facilitating the growth of data archiving.

Introduction
As the central bank of the United States of America, the Federal Reserve System (FRS) consumes vast quantities of economic, financial, and organization structural data. These data are used for making monetary policy, conducting banking supervision, performing economic research, and implementing consumer protection policies. The Federal Reserve System is comprised of the Federal Reserve Board of Governors (the Board) and twelve regional Federal Reserve Banks (Reserve Banks). The focus of this paper is on micro data archived at the Board. However, much of the data archived at the Board is supplied by and used by the Reserve Banks as well as Board staff.

Both micro and macro data are housed at the Board. Micro data refers to institution level data whereas macro data refers to sector, industry, or economy-wide aggregated data. Industrial Production, which is a principal indicator of economic activity in the United States' industrial sector, is a good example of the Board's use of micro and macro data and how they interrelate. The Board receives input to the industrial production indexes from a variety of sources including sample surveys of independent firms, trade organizations, and other agencies. The micro data are weighted and aggregated to generate the macro data series.

Micro Data Collection, Editing, and Storage
The Board acquires micro data through a variety of methods. The Board purchases data from independent vendors for information on sectors such as commercial interest rates, stock and bond market prices, and nonfinancial industries. The Board also receives some micro data from other regulatory agencies. Finally, the Board conducts surveys to collect data within the financial services industry. Over 60 surveys of varying frequency are currently collected. Some surveys consist of a sample of entities from a population (such as commercial banks). The population is then estimated from the sample and the macro data are produced. Relatively few surveys are a periodic census of a population such as the Consolidated Report of Condition and Income for a Bank, commonly known as the Call Report.

The Federal Reserve System maintains an ongoing census of the banking industry's structure by collecting data on structure changes either directly from the bank, the bank holding company, or from the bank's primary regulator. The Federal Reserve System's structure system contains descriptive, geographic, and ownership information. It also identifies events such as mergers between depository institutions and bank holding companies. The structure system includes all bank holding companies, all banks and their branches, all thrifts, and all credit unions in the United States. It also contains some foreign bank and other financial and economic sector data.

Because so much micro data are collected directly from businesses, the Federal Reserve System has a large infrastructure to collect, edit, process, and distribute data. Although the process varies slightly among surveys, the majority of surveys follow the process outlined in Figure 1. The collection process begins with the reporter (usually a financial institution) transmitting the data requested on a form to the responsible Federal Reserve Bank. The transmission processes range from reporters mailing or faxing paper forms to secured electronic data transfers that load the data directly into the Board's editing system, depending on the survey and reporter.

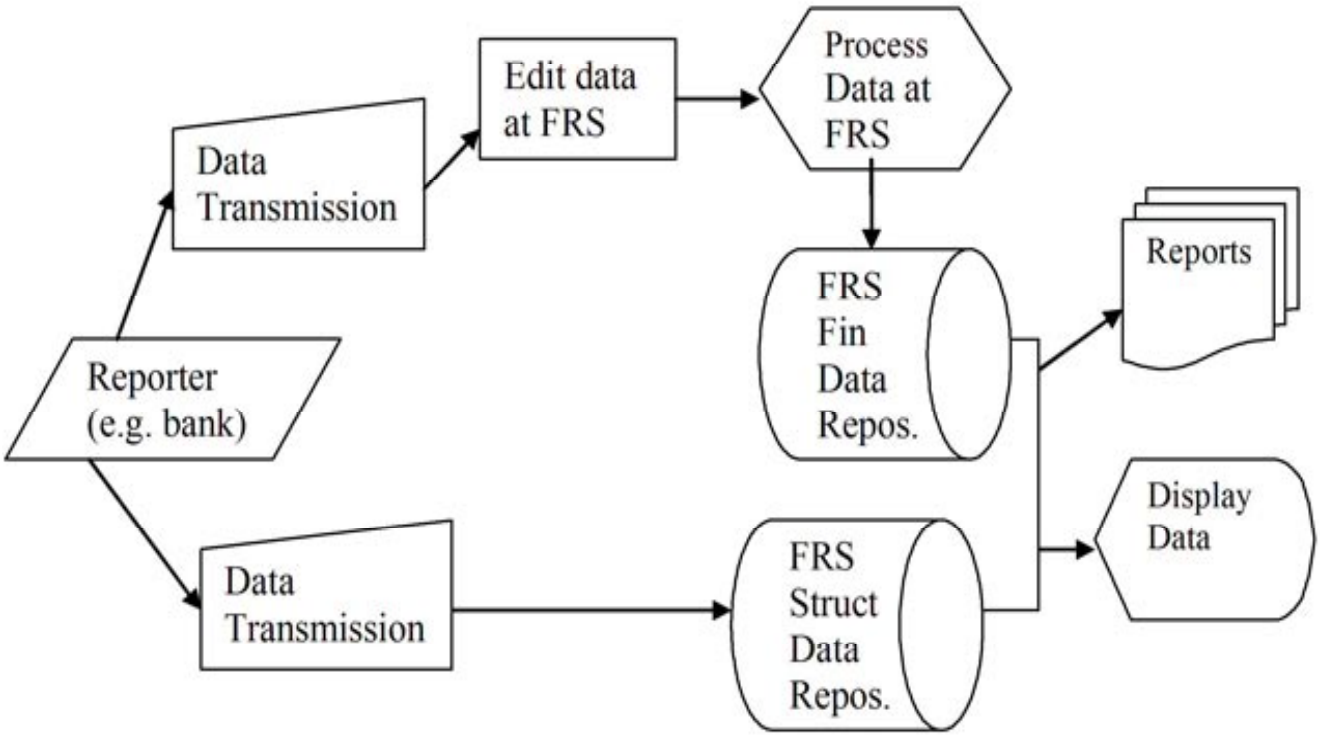

Figure 1

The editing system is a data repository designed to quickly and accurately edit massive[2] quantities of data. The system is parameter-driven and new data surveys and edits can be added relatively quickly. Four types of edits are performed in the editing system; validity, quality, intraseries, and interseries. Validity and quality edits confirm that the data within a transmission are consistent and do not violate any standard rules. For example, validity edits ensure that all required fields are not null and that impossible scenarios (such as a negative number of employees) do not occur. Quality edits compare the data within the transmission to ensure that the data follow accounting (or other) rules and generally make sense. For example, an accounting rule for a balance sheet is that assets = liabilities + owner's equity. Since rounding can create some slight inequalities tolerances are usually associated with each edit. The tolerances can include percentage differences and/or unit differences. Intraseries edits look for consistency between periods and over time. For example, if a bank reported a 1000 percent increase in total assets over one quarter it is likely that extra zeros were erroneously added to the submission and the data should be reviewed for accuracy. Intraseries edits also have percent and unit tolerances associated with each edit.

Since there are some reports that collect similar data but on different frequencies interseries edits are also performed. These are edits that compare similar data items on different surveys for comparable dates. For example, the bank's weekly report of deposits collects some information that is similar to the quarterly Call Report which contains some deposit information. The deposit information on the two surveys can be compared to ensure that they are comparable.

Once the data have been edited they can be stored in the Financial Data Repository (FDR). However, for some surveys there is additional processing that is done to the data such as aggregating data, deriving commonly used data items, and incorporating additional data already housed within the Federal Reserve System. After the processing is complete the data are stored in the FDR DB2 relational database.

The banking industry structure data are processed similarly but in a different collection and editing facility because the nature of processing and editing structure data is different from that of financial data. The structure data repository is also in DB2 but the collection, transmission, and editing are done on Microsoft SQL servers. The structure data are broken into different databases to capture attribute (name, location, entity type), relationship (parent, ownership), and merger (acquirer, type of merger) information. The edits

in this system ensure consistency within the databases and adherence to a predefined set of rules for displaying the data. One example of an edit is that the attribute data for an entity must end if the entity is acquired in a merger.

Micro data are acquired through a variety of additional methods to meet the special needs of individual surveys. These other methods include contracting with research organizations, direct mailings of surveys, purchasing commercial products, and supplying custom software. The method of data collection and processing used depends on the type of data, frequency, population being surveyed, and other special needs of the survey. How these data are stored and accessed also varies depending on the method of acquisition, user needs, and license restrictions.

Micro Data Documentation

Because the central bank collects data ranging from bank balance sheet values to kilowatt hours generated, the need for centralized and comprehensive documentation became apparent in the early years of data archiving. For the surveys collected and stored via the method described in Figure 1, there is an on-line dictionary that defines the characteristics and content of each of the surveys. This dictionary is called the Micro Data Reference Manual (MDRM) and contains descriptions of the surveys as well as meta data for each data item stored. The data are organized by survey and consist primarily of financial and structure data. The MDRM documents the labels and values (meta data) associated with each data item in a survey. A web interface is used to access and display the MDRM meta data.

For the collection and storage process each survey is given a mnemonic such as EDDS (the report of deposits). Each accounting concept or data item collected is given a number such as 2200 (total deposits). Because comparable accounting concepts are collected across various surveys the same number is used for all comparable data items regardless of the survey. Combining the survey mnemonic and variable number references a specific data item within a specific survey. The combined mnemonic and number is commonly referred to as the MDRM number. For example, the MDRM number SVGL2170 refers to total assets on the Thrift Report of Condition and CUSA2170 refers to total assets on the Credit Union Report of Condition.

Within some surveys the same accounting concept may be collected several times but for different populations or periods. In these cases, multiple mnemonics (a.k.a subseries mnemonics) can be used for one survey. An example of this is in the EDDS survey, total deposits are collected weekly for each day of the week. To identify which day of the week the data apply, the EDDS mnemonic is broken into EDD1 to represent Tuesday's data, EDD2 to represent Wednesday's data … EDD7 to represent Monday's data.

The MDRM has three main components; the reporting forms, the mnemonics information, and the data dictionary. The reporting forms section is a historical PDF library of all the forms and instructions used to collect micro data. The forms are the visual representation of what data should be reported and the instructions are provided to give detailed information about who, how, when, and exactly what to report. The mnemonics section describes each survey, provides a hierarchy of subseries mnemonics, and links the mnemonics to the published reporting form names.

The data dictionary is the heart of the MDRM and defines each data item collected on any of the surveys. For each data item the MDRM provides the starting and ending dates it was collected for each survey. It also provides a long caption, a short caption (similar to the long caption but limited to 40 characters), a confidential indicator, a data type (financial, structure, ratio, or derived), and a long description which provides the detailed instructions, history, and idiosyncrasies of an accounting concept between surveys. The web interface displays all surveys associated with a specific data item number or all data items within a survey. Each survey also has a glossary which identifies unique information pertinent to the survey such as EDD1 represents Tuesday data.

The documentation of data purchased from vendors or collected through contractors varies between surveys and sources of the data. Vendor commercial packages generally have user guides but there is not currently a central documentation facility.

Macro Data Collection, Storage, and Documentation

Most of the economic macro data used at the Board is housed in FAME databases which are designed to store large volumes of time series data. As with the micro data, the macro data are collected from a variety of sources in a variety of formats ranging from PDF to delimited files. The sources include Board staff's aggregations of micro data, other government agencies, research organizations or universities, private organizations, and commercial data providers.

Once the data are received they are processed and renamed, using a Board nomenclature, and stored in a temporary database. While in the temporary database, several quality edits are run against the data. The edits look for nulls, excessive variability over time, and noncontiguous date problems. The edit routines run against the data are standard for all series with a few minor exceptions. Once the quality of the data is confirmed the series are loaded to the production database.

The macro data used for most official forecasting are stored using a hierarchical nomenclature in FAME databases. The main FAME databases hold either US or international data.

To get to a specific series, users can 'drill down' through the nomenclature to retrieve the desired time series. For example, the US database includes income data for which the first letter of the nomenclature is 'Y'. Within income there are several categories including personal ('P') which also contains several categories. Within personal income is disposable outlays ('D.O') which has the full nomenclature of Y.P.D.O. These data can be viewed at this level or they can be further disaggregated and viewed. In the international database the country is denoted by a suffix, such as .UK.

The FAME database allows for some self-documentation and each series has several attributes that provide descriptive data. The attributes include information regarding where the data originated, where the data are stored in FAME, Board contacts, periodicity, adjustments to the data, units, unit multiplier, currency, update frequency, and several attributes that describe formulas. The nomenclature shows the relationship to parent series.

Because hundreds of thousands of macro data series are stored in FAME, there is also the need for data documentation. To aid in the documentation and navigation of macro data several web-based tools are available at the Board including a source book of nonfinancial economic data. The source book lists the various sources of data, what data are received from each source, data definitions, the data collection process, and adjustments to the data, as well as information specific to the source.

Challenges and the Evolution of Data Archiving

As the need for more and varied data grows, we have encountered new challenges in the collection, storage, access, and documentation of data. Twenty years ago, the majority of micro data used at the Board were collected and stored in a model similar to Figure 1. Today, the reliance on market information and data obtained through other sources eclipses the volume of collected data, forcing users to spend more time researching what data are available and learning how to access the data.

Catalogue

One of the greatest challenges is the cataloguing of all data purchased at the Board. As in academia, there are often several individuals independently studying different aspects of the same industries. These different studies can be performed in unrelated areas of the Board but have similar data needs. For example, when Citicorp Bank Holding Company and Travelers Insurance Company merged, the supervision community was interested in insurance company micro data to determine the effect of the merger on the bank holding company's safety and soundness and the banking industry in general. Simultaneously, the economic research departments are interested in the insurance industry's impact on the economy and financial markets.

As with most corporate, academic, or government bureaucracies, budgets and responsibilities are dispersed throughout the agency. As the volume of data purchased has increased, ensuring that similar data are not purchased multiple times has become increasingly difficult. To address this challenge, the Board has created a "Data and News Catalogue" of purchased data. The catalogue contains information for each data purchase including the database, the vendor, internal contacts, the form of access, and license information. It also contains a brief description of the data purchased and links to the vendor's website or the database if available.

The catalogue was originally populated by surveying the budget areas to identify all data expenditures. The catalogue is maintained by soliciting information from the budget areas as well as by having some areas update it directly. An annual review during the budget cycle catches additions or deletions not otherwise captured during the year.

The Board also maintains several ongoing databases, such as merger adjusted balance sheet data for banks. The time and effort to create and maintain these databases are extensive. To avoid possible duplication of effort or purchasing of data already captured in-house, the catalogue also describes ongoing databases created and maintained by Board staff. Once users know what data are available, the next step is accessing the data.

Storage and Cross Referencing

Much of the purchased data are stored within vendor software and can only be accessed via the vendor software. Oftentimes the data can be exported to a SAS dataset or Excel file that can be further manipulated and analyzed or combined with other data. Many license agreements limit the way data can be stored and the usability of vendor software limits the ability to standardize the storage and documentation of micro data. The volume and complexity of purchased data also make it impractical to try to put all micro data in a standard storage repository with documentation such as the MDRM or FAME nomenclature. The uniformity of micro data storage is just beginning to be pursued but is hampered by license agreements and data availability. Similar licensing problems are encountered on the macro data level and in some cases licenses need to be negotiated to allow for the loading of data into FAME.

Data consistency over time is another challenge. As users' needs and the economy change, so do the data. For example, the calculation of goodwill has changed over the last decade as accounting rules and regulatory requirements have changed. In addition to changes over time, there are often different names for the same accounting concepts across industries such as capital vs. net worth. To address this challenge, wherever possible, micro data items with like meanings are given the same numbers in the MDRM

or new data items that can be used over time are derived from the existing data. In addition, slight changes to micro data items that result from changes in the markets or regulatory changes are generally captured through notes to the description rather than giving the item a new number. Changes that have a large immediate impact require new numbers.

For macro data, the FAME nomenclature enables historical series to be link with current series. For example, when the U.S. Census Bureau changed from using Standardized Industrial Classifications (SIC) to North American Industrial Codes (NAIC) there was a one-to-one relationship between a number of the SIC and NAIC codes. Therefore, series for a SIC code that had a direct relationship to a NAIC code were flagged as historical series to the NAIC series.

Researchers also have increasing needs for more data. This is particularly prevalent in the macro data series. Specifically, there is demand for geocoded data including an increasing demand for geocoded housing data. For example, there is a desire to be able to drill down from US data (1 series) for a specific series to specific regions (4-12 series), states (50 series), MSAs, and counties. This type of geocoding on a large number of series can have an exponential effect on the volume of data stored.

In addition to needing more data, researchers have an increasing need for cross sectional data. As technology improves and sources of data grow, the availability of market data has increased. Purchasing data from a vendor also adds flexibility to a research project because the lead time is short and there isn't an investment in infrastructure, so changing the sources of data is easy. In recent years, evaluating stock and bond market data in conjunction with corporate structure and regulatory accounting data has become increasingly feasible. These different types of data can come in a variety of formats and from various sources. Linking the data between sources is challenging because it is difficult to join data from different vendors or in-house databases. To address this challenge, the Board is currently evaluating compiling a database of key fields. The Board's entity identifier (ID_RSSD), the stock market ticker, and the bond CUSIP number are some of the key fields being included in the date-sensitive database. Once complete, this database will allow users to join data from different sources based on the key identifiers and date of the data.

Technology Advances

The introduction of eXtensible Markup Language (XML) is allowing the data archiving process to evolve in new ways. XML brings the ability to tag individual concepts (text or numbers) with context so data can be searched quickly and accurately. For example, the word 'bank' refers to a financial institution or the side of a river. XML enables each word to be given a different tag that will signal to a search engine which meaning the word has in the context in which is it being used.[3]

XML is valuable as a transmission protocol because it allows for sending and receiving large quantities of data. To further facilitate the transmission of data, several groups are developing transmission standards. The accounting industry is focusing on the XBRL standard which is designed for financial statement data. Another transmission standard for statistical data is the Statistical Data and Metadata Exchange (SDMX). The Board is currently implementing production systems using each of these transmission protocols.

XBRL is being used for an interagency project between the three primary U.S. bank regulators to collect Bank Call Report data over the internet. SDMX is being used for the downloading of macro data in a new bulk data download facility available on the Board's website.

Conclusion

The international and domestic economies, financial markets, and economic models continue to become more complex requiring more data to evaluate the growing complexities. The increasing number of sources of data and the increasing volume of data add to the challenges faced by data librarians and other data archivists. Advances in technology, such as XML, help to facilitate the resolution of some of these challenges but at the same time create additional technology hurdles. To address the challenges that lie ahead, data archivists need to ensure that they have organized, flexible, and well-documented data archives.

Meta data, naming conventions and nomenclatures, and other data documentation are becoming more important to describe large and diverse data archives. Data quality verification and definitional consistency are necessary to ensure the usability of the data archived. Maintaining meta data and complying with nomenclature rules are time-consuming and tedious tasks, but if you have a large group of users over a long period it will save resources and avoid frustration. Finally, centralizing data of various file types can be difficult and time-consuming but ensures that data are not lost and reduces the burden on end users needing to know how to use a variety of data access tools.

Bibliography

Bayard, Kimberly and Morin, Norman. "Industrial Production and Capacity Utilization: The 2003 Annual Revision." Federal Reserve Bulletin Winter 2004. 20 December 2004. http://www.federalreserve.gov/pubs/bulletin/2004/winter04_ip.pdf.

Cannon, Sandra, Chief of Economic Information Management. Federal Reserve Board of Governors. D.C. Personal Interview. 3 January 2005.

FAME Time Series Database. 2003. Sungard Data

Management Solutions. 20 December 2004.

http://www.data.sungard.com/infrastructure/fame/index.htm

Federal Reserve Statistical Release G.17 Industrial Production and Capacity Utilization. 14 December 2004. Board of Governors of the Federal Reserve System. 20 December 2004.

http://www.federalreserve.gov/releases/G17/Current/default.htm

Wallison, Peter J. "Enhanced Business Reporting Gets a Start." The New Republic Magazine, December 20, 2004, pp ON/1 – ON/4

Endnotes

[1] Linda F. Powell, Board of Governors of the Federal Reserve System, Washington, D.C. Email: Linda.Powell@frb.gov.

[2] Data from over 2,500 bank holding companies and over 18,000 depository institutions, each of which reports hundreds of financial statement data items, are processed and stored at the Board each quarter. This is in addition to other weekly, quarterly, and annual reports that follow the process outlined in Figure 1.

[3] P. J. Wallison, "Enhanced Business Reporting Gets a Start," The New Republic, 20 December 2004, p. ON/3