# Everything but the Kitchen Sink: Building a Metadata Repository for Time Series Data at the Federal Reserve Board

*by San Cannon and Meredith Krug**

In support of the Board's duty to conduct monetary policy for the United States, the research divisions at the Federal Reserve Board use a variety of time series data for both research and forecasting. The collection, maintenance, and upkeep of more than 50,000 time series from more than sixty sources in a central location are daunting tasks; documenting the metadata for the compilation and use of these data is even more so. We are currently building a comprehensive metadata repository that links three kinds of metadata about our time series: structural metadata, reference metadata, and operational metadata. Many of the pieces of this puzzle currently exist in an array of disparate formats: attributes in a proprietary database, HTML pages on a website, Word documents buried on a file server, etc. We are bringing these pieces of information together in a relational database setting to allow users to search and display relevant metadata for a particular series or economic concept. In addition, we are working to make the metadata entries in the repository time sensitive to accommodate the database of contemporaneous "real time" or "snapshot" time series that we are building for future research purposes. This paper will examine the different types of metadata gathered for time series data, and how this information is collected, stored, and made available to staff at the Federal Reserve Board.

**Metadata Types**

The concept of metadata or "data about data" raises questions concerning what users need to know about different pieces of information. We have identified three types of metadata that cover the information we gather. The first and most fundamental type is **structural metadata**. Structural metadata are usually a small set of concrete details that identify and define a particular time series; they can often be expressed as short strings of text that convey the information essential to the meaning of a number. For example, users would have difficulty interpreting the statement that output was 12487.1. What kind of output? How is it measured? When was it measured? These questions may seem basic, but their answers are integral to the correct interpretation of this string of digits as economic information. If a few short strings such as description, country, time period, and unit of measure were defined, and you were told that the gross domestic product of the United States for 2005 was 12,487.1 billion U.S. dollars, the number would now have meaning.

Further details on the construction of a statistic are better classified as **reference metadata**[1] Reference metadata contain more detail about the calculations behind a number and may describe the data collection, sampling methodology, preparation of the estimates from the sample data, treatment of revisions, and the reliability of the estimates. Reference metadata are often presented as a document provided by an issuing agency in an article or bulletin, which may eventually appear on its website. Researchers often need to understand the more complicated details of how a series is constructed. Reference metadata are invaluable when deciding on the appropriate series to use in research, especially if more than one possible measure is available. For example, the Department of Labor publishes the Employment Cost Index (ECI) and the Employer Cost for Employee Compensation (ECEC), both of which have data on the cost of labor to employers. Typical structural metadata on these statistics would indicate that the ECI series are indexes with a base year of 2005 = 100, and that the ECEC series are measured in current U.S. dollars. The reference metadata, however, would get to the heart of the calculation behind the time series and show that the ECI indexes use "fixed weights to control for shifts among occupations and industries" whereas compensation figures from the ECEC use "current weights to reflect today's labor force composition." A researcher can then decide which construction best represents the economic concept under study.

The third type of metadata we collect can be classified as **operational metadata**. These are the processing instructions that explain exactly what is done with the data once they are received and exactly how it is accomplished. This type of metadata is extremely valuable to the data maintenance staff and facilitates the job of collecting, uploading, and disseminating information. Typical operational metadata contain information on the source agency's publication schedule and procedures, file formats of the data, the location of updating programs and how they work, archiving information, data contact information, and other special considerations.

All three types of metadata are closely connected. A change in the reference metadata could have consequences for the structural and operational documentation. For example, when the Census Bureau changed the sampling universe of permit-issuing places from 19,000 to 20,000 for the new residential construction statistics, the reference metadata explain in detail the change in definition. The structural metadata, however, would need to note that there is an important break in the continuity of the series, and the operational metadata would need to be updated with the operational consequences of processing the newly defined data.

**Where We Started**

The two main production databases used by the research departments of the Federal Reserve Board contain about 62,000 time series (52,000 domestic and 10,000 international). These numbers are inflated by multiple versions of series in different frequencies (monthly, quarterly, etc). Removing the frequency dimension, there are about 45,000 unique economic measures in the two databases. Our current "method" of storing all three types of metadata for this information is a hodge-podge of legacy systems that differ by metadata type. In the past, different information was stored in different formats on different platforms, depending on where the metadata were originally placed, who placed them there, and who needed access to them.

**Structural Metadata**

The structural metadata for our time series data are stored with the data in a FAME database. Designed to manage times series, FAME software allocates some characteristics when a series is created, such as name, description, database name, first value, and last value. While these native attributes are useful, they do not provide enough information for the data to be uniquely identified and used correctly. So that the data will be more useful to the economists and analysts who use them, we have implemented two additional metadata constructs to convey information about the series: a detailed, hierarchical nomenclature system that restricts the name of the series, and a set of additional attributes to help capture some of the more critical information about each data series.

The hierarchical nomenclature system is our key to storing, in a single repository, a vast collection of data series covering a wide range of topics. The system was developed in the 1980s from a skeletal structure purchased from the consulting firm of Townsend Greenspan, and over the past quarter century Board staff have expanded its scope. It allows for the categorization of a variety of macroeconomic concepts and is adaptable enough to accommodate the changing economic landscape. We construct our series names from prefixes, roots, and modifiers with a frequency designation, but only the root and the frequency code are required.

There are 62 high-level roots covering broad economic topics such as gross domestic product, consumer expenditures, employment, and price indexes. Each of these high-level keys has a "branch" of subcategories with further breakdowns as necessary. The result is 22,000 unique "roots" describing the economic concepts. Some roots are succinct: For example, PJP indicates producer price index, where PJ is the high-level key indicating the broad concept of price index and the additional P indicates the subcategory of producer prices. Such a breakdown might be depicted as:

```
PJ: Price index
PJ.P: Producer
```

Other roots, such as the one for two-year installment loans to consumers (RIFLPBCIPLM24), are painfully detailed:

```
R: Rate
R.I: Rate of interest in money and
capital markets
R.I.F: Federal Reserve System
R.I.F.L: Long-term or capital market
R.I.F.L.P: Private securities
R.I.F.L.P.BC: Commercial banks
R.I.F.L.P.BC.I: Consumer installment
loans
R.I.F.L.P.BC.I.PL: Personal loans
R.I.F.L.P.BC.I.PL.M24: 24-month loan
```

Once the root is established, a series name can have up to four "modifiers"—special codes that add more information and are preceded by an underscore. Modifiers describe concepts such as base year, data source, product designation, industry (NAICS and SIC definitions), region of the United States, country, currency, import designation, export designation, commodity group, occupation (SOC categories), asset class, age group, and market category[2]. Modifiers can also be used to identify a series as seasonally adjusted or not seasonally adjusted, break adjusted or merger adjusted or to indicate that a series contains seasonal factors. In addition, we can identify series that are simple calculations from other series, such as different types of percentage change. A straightforward example with only two modifiers is the monthly Producer Price Index for Electric Power. As shown here, it has two modifiers listed after the root: one indicating the Bureau of Labor Statistics commodity code and one indicating that the series is not seasonally adjusted:

```
PJP_G054_N.M:
PJ: Price index
PJ.P: Producer
_G: BLS commodity group
05: Fuels and related products and
power
```

```
05.4: Electric power
_N: Not seasonally adjusted
M: Monthly
```

Finally, a series must have a frequency code at the end of the series name, preceded by a dot. This Producer Price Index ends with ".M," indicating that it is a monthly series. Other examples of frequency codes are .A (annual), .Q (quarterly), .B (business daily), and .WF (weekly Friday).

The meticulous identification of the economic concept with our nomenclature, however, is not sufficient structural information for our researchers and analysts. Our second construct is a set of additional attributes stored in the proprietary database format with the data. These attributes provide needed structural metadata not included with the native attributes allotted by FAME when a series is created.[3] Currently there are fifteen additional attribute categories. Some attributes are administrative, such as when the series was updated (to the second), who has permission to make changes, what group is responsible for the series, and who to contact with questions. Other attributes are more substantive, such as strings containing the source agency and publication, the table and line number on which the data are presented, restricted value attributes for units, unit multiplier and currency, an annual rate Boolean, and the frequency of updates. For example, the additional attributes for the Producer Price Index for Electric Power would be

```
AGENCY:          Department of Labor
/ Bureau of Labor Statistics
PUBLICATION:     Producer Price Index
release
TABLE:           PPI release table 3
LINE_NUMBER:     39
UNITS:           Index: 1982 = 100
UNIT_MULT:       One
CURRENCY:
ANNUAL_RATE:     NO
SECTION:         EIM
DATA_GROUP:      PPI:PRICE
CONTACT:         San Cannon
UPDATERS:        Meredith Krug, Other
EIM updaters
UPDATE_SERIES:   PJP_G054_N.M
UPDATE_FREQ      Monthly
TIME_STAMP:      20060418083440.00
```

**Reference Metadata**

Many times knowing the nomenclature or the discrete list of attributes for a series is not enough information to make a decision about the appropriateness of a series for a particular use. For detailed methodology or reference metadata, economists in the research divisions can refer to the source agency's documentation. In addition, we maintain a Data Sourcebook—a set of web pages created from the documentation compiled by a research economist on staff. This information, gathered mainly from the source agency's information and from conversations with staff members at different agencies, started life as word processing documents and were then converted to static HTML pages. These pages outline the more detailed methodological information for most of the major statistical releases from which we gather data. Typical topics covered include the source, useful links, principal data provided, history, concepts and definitions, sample design, data collection, revision information, reliability information, seasonal adjustment details, and typical uses.

For some statistical releases, the metadata are fairly straightforward. The entry for new residential sales data, for example, has only three definitions of interest: "sale," "houses for sale," and "sales price." Other series require as many as ten concepts. For example, the sourcebook entry for the Current Population Survey, from which we obtain the household estimates for the Employment Situation release, explains the meanings behind "civilian noninstitutional population," "employment," "unemployment," "civilian labor force," "unemployment rate," "duration of unemployment," "reason for unemployment," "not in the labor force," and "discouraged workers."

These details are important for distinguishing between various definitions of economic concepts and for understanding the properties of some series. In these detailed pages economists will find notations, for example, about the difference between "all persons" and "all employees" measures for productivity and cost data and about the types of adjustments the Bureau of Labor Statistics makes for that distinction in the payroll data.

**Operational Metadata**

Clean structural and reference metadata are vital to a data user's understanding of a time series, but more information is required for the data maintenance staff to properly maintain the actual time series. We have two operational metadata tools to keep track of the statistical releases that need to be updated to the production databases and outline the procedures for updating them. The first and most essential is our Data Release Calendar, which contains all the statistical releases that we follow and keeps track of exactly what date they will be published—daily, weekly, monthly, quarterly, or annually. Like the Data Sourcebook, this calendar was converted to static HTML from a word processing document.

Our second tool, the Data Documentation System, stores the procedures that we follow for each statistical release. Each page was originally created as a static document in Word and HTML. The files varied in structure and organization, depending on who initially wrote them. A few pieces of information were common across pages,

such as who had updating responsibility for the data and some external contact information. Most content, however, varied according to what each author thought was important. Some documentation had secondary pages with additional information, such as screen captures and cryptic instructions to "hit Control-P" twice. The only consistent element was a loose description of how to incorporate the data into the database on publication day, but even that was frequently out of date.

## Building a New System

The three types of metadata are of interest to different sets of users: structural metadata are important to all the users of a time series; reference metadata are important to the smaller group of analysts doing more detailed research; operational metadata are important only to the people managing the data. As the audience size for each type of metadata slowly decreases, our influence over the metadata as data managers increases. We have little say in what structural metadata are specified; the decisions are usually made by committee, and programming for the additional attributes is done by other staff members. We have slightly more to say about the reference metadata, as the decisions are usually made by the analyst or economist most familiar with the data. Finally, we have complete control over the specification, collection, and storage of the operational metadata for the data we maintain.

### Operational Metadata

In working on a system to tie all three types of metadata together, we started with the operational metadata. Several years ago, we moved our first operational tool, the Data Release Calendar, from static HTML pages to a relational database. Each statistical release is its own entry in the calendar and provides links to an interface for editing the entry. As publication dates are known, releases can be easily added to, moved within, or deleted from the calendar. The interface also provides information such as the individual responsible for that update; a link to the statistical release on the source agency's website; and a link to the relevant page in our second tool, the Data Documentation System.

After moving the calendar information to the relational database, we improved the data documentation. First, we identified the key operational metadata categories for maintaining our data that apply to all statistical releases. Then we stored each release and its standardized information in the relational database and created a web interface. The main data documentation page is an alphabetical listing of all the statistical releases from which we retrieve data—currently around 100 releases and growing steadily. Linked to each release entry is a secondary page containing all the operational metadata categories from the database. These categories explain how we obtain all the information we need and what we do with the information once we have it. The fields include publication schedule, the releasing agency and contact information, the location of the press release, and how to retrieve the data. We also detail how our update programs work, what other programs are run after the data are updated, and who is notified of their availability. We maintain additional documentation about what happens when there is a revision or when our usual update methods are unavailable. An additional section of notes allows the user to document important issues, lessons learned, or information gathered in phone conversations with agency contacts.

This new Data Documentation System is similar to a "restricted wiki," allowing only the data maintainers to update it and providing simple formatting options for different types of entries. For each statistical release, an edit button allows a user to edit the text fields and make selections from drop-down boxes and radio buttons for each defined category. Behind the interface is an SQL database table with rows indicating the statistical data product and columns indicating the various fields. Editing the appropriate field on the appropriate page is a simple task that updates the database so that the new information can be rendered on the presentation page immediately.

The transition to this system from the Word document-turned-static HTML page was lengthy and painful but the payoffs have been tremendous. We now have a clear, concise, and consistent framework in which to store our operational metadata, which is easy to edit in a timely manner. Less energy is expended in deciphering and maintaining the documentation. The new system also provides an organized way for all data maintainers to communicate with each other and has made training new data managers much easier.

### Reference Metadata

Our next step is to work on the other type of metadata over which we have some influence. We are currently planning to alleviate some of the pain of editing HTML code by hand by transferring the Data Sourcebook from static HTML pages to a table in the SQL database.

This system is in its very early stages but is similar in structure to the Data Documentation System. On the main page, each row is a statistical release and has an edit button to enable the "restricted wiki" interface that accesses the database table containing all the reference metadata. Each entry links to a secondary page with standardized categories for the reference metadata with text fields that can be updated. Currently, the fields we have identified are: data source, principal data provided, concepts and definitions, history, data collection, preparation of estimates, revisions, reliability, seasonal adjustments, uses, secondary uses, additional resources, miscellaneous, and references.

There is the possibility for much more variation in the

reference metadata categories than in the operational metadata categories. For example, every statistical release will have a source, but, as noted above, the number of concepts and definitions will vary. We are still working on the best way to handle such variations. Regardless, our hope is that the relevant analysts and the data maintenance staff will be able to easily update the information in the system in a simple and timely fashion.

**Structural Metadata**
Because the structural metadata are so closely tied to the actual data, it is more practical to have the primary storage location in the database with the time series. To complete our relational database repository, we simply copy the structural information to a table in the SQL database. Those who are not interested in using the data in FAME often want to know how to find what they want and get it out of FAME. We have built a web interface with a search facility that researchers can use to find time series in the database based on the structural metadata stored in the relational database. They are then able to extract the data from the database and use it somewhere else.

## The Challenge of Time

The next step is to incorporate the element of time into our integrated metadata. Neither data nor metadata are static, and changes to the former should be documented by changes to the latter. We are currently building a library of "vintage" or "real-time" data. Similar efforts are underway at the Federal Reserve Banks of St. Louis and Philadelphia, but instead of re-creating history as those projects are doing, we are building a collection for future use.[4] Every day we store a snapshot of the production databases in their native format. So far we have stored about one year of databases, but in a few years we will be able to look back and identify exactly what information the research staff had for any given business day.

Capturing snapshots of the data over a period of years, however, may not allow for a complete understanding of the changes to those data over time. The structural metadata stored in the database with the time series should reflect the correct contemporary information, but for many purposes that information will not be enough. Changes to the components of an aggregate series may not be reflected in the mnemonic or in any of the stored FAME attributes. If the source agency changes their method of seasonally adjusting data or redefines the population it is sampling, the changes would be noted in the reference metadata but not in the structural metadata. To understand the full meaning of a time series, it is essential to know the precise definitions and data construction for a given time period.

For full use of the rich metadata resource we are working toward, the metadata storage must be able to depict contemporary information. To capture it, we need to add time-sensitivity to the relational database. We need to be able to capture when edits to various fields in the metadata are made so that we will know the period for which the information in that edit was valid. Then researchers will be able to see the reference metadata as they were on any particular date and thus make the most of the vintage data available to them.

As before, we are starting with the operational metadata storage and will then work on the reference metadata storage. We are currently building a time component into our Data Documentation System. We maintain a detailed archive in which data files and programs are automatically stored every time a data file is processed. The archive enables the data managers to restore data in case of a technical problem or to track down the source of an error. Being able to reproduce the data-updating instructions for any publication date allows us to take full advantage of the archive we have built; we will have complete instructions on what to do with the various files that are stored with date and time stamps. Next, we plan to work on adding a time component to our Data Sourcebook.

**Final Thoughts**
The goal for this project is to have a gateway to data and metadata that is informative, simple to use, and time sensitive. The biggest challenges we face are those of information retrieval and interface design. We may have a wealth of useful information in our database, but if users are presented with too much information in an incoherent way, the collection is worthless. We need to build a search facility that will allow users to find the information they need for a given point in time, regardless of the type of metadata. After the pertinent information has been retrieved, we need to present it clearly and concisely to the user. Once these hurdles have been cleared, the metadata repository for the research staff at the Federal Reserve Board will be an invaluable research tool.

* Paper presented in the session "The Essential Role of Metadata in Resource Discovery" at the IASSIST 2006 in Ann Arbor by San Cannon and Meredith Krug, Federal Reserve Board. Contact: San Cannon, scannon@frb.gov. Chief, Economic Information Management, Federal Reserve Board, Washington DC 20551.

**Endnotes**
1 The distinction between structural and reference metadata, as well as the terminology, was adopted from the SDMX Initiative's Metadata Common Vocabulary. See Statistical Data and Metadata Exchange Initiative (draft, March 2006), "SDMX Content Oriented Guidelines: Metadata Common Vocabulary," SDMX, www.sdmx.org/news/document.aspx?id=146&nid=67 (accessed June 26, 2006).

2 Details on the North American Industrial Classification System (NAICS) and its relationship to the Standard Industrial Classification (SIC) system can be found on the Census Bureau website at www.census.gov/epcd/www/naics.html. Details on the Standard Occupational Classification (SOC) system can be found on the Bureau of Labor Statistics website at www.bls.gov/soc/soc_majo.htm.

3 The "native" FAME attributes include information about the object type; measurement information that affects frequency conversions; index information including type of index and first and last value indicators; and creation and update dates.

4 For information on the Saint Louis project, see Katrina Steirholz (2005), "Economic Data as Snapshots in Time," IASSIST Quarterly, vol. 29 (2) (Summer), p. 5. To access data and documentation from the Philadelphia system, see the Federal Reserve Bank of Philadelphia's website at www.philadelphiafed.org/econ/forecast/readow.html.