Data and Knowledge Management at the Federal Reserve Board

Andy Boettcher[*1]

Abstract
The Federal Reserve Board (Board) purchases and creates numerous datasets to support its role in monetary policy, banking regulation, and consumer protection. To better manage these datasets, the Board has built a metadata repository called the Data And News CataloguE (DANCE), which stores descriptive dataset characteristics. The growing number of datasets and their corresponding security and licensing intricacies motivated a data initiative in which the Board's research community identified enhancements to DANCE. Planned improvements include: the addition of Dublin-Core standard metadata, the communication of changes in metadata, and the dissemination of metadata on new datasets. The improvements are expected to enhance collaboration between research units of the Board, which will in turn enable better research. This paper will chronicle DANCE's original role within the organization and its transformation into a knowledge management solution.

Background on Data Management
The growing interaction between banking, the financial economy, and the real economy has necessitated cross-departmental projects within the Board and the Federal Reserve System. For example, three different research departments conduct market discipline[2] research using similar data. As the number of multi-department research projects increase, data management practices are evolving to facilitate these changes. Historically, data management practices were relatively fractured[3] across the Board as well as the Federal Reserve System[4]. Research departments that purchased or created datasets tended to silo the data, documentation, expertise, or any combination of the three. If research projects were always intra-departmental, then this approach to data management would suffice.

The acquisition of data also presented challenges. To purchase a dataset, a department must confirm that the Board does not already have access to a similar dataset. Since the Board lacked a central metadata facility, attempting to find similarities across undocumented datasets proved particularly frustrating, time-consuming, and inadequate. Arguably, more resources were spent determining if similar datasets existed than the cost of purchasing a duplicate dataset.

DANCE History
DANCE was first conceived in 2002 as a tool to encourage using newly purchased datasets to Board researchers outside the purchasing group. Increasingly, research staff's forecasting and working papers required purchased datasets, yet knowledge of these datasets remained primarily with the purchasing group, not the larger user group. Therefore, DANCE's foremost purpose was to remove data silos by serving as a search hub linking locally documented and stored datasets with the larger Board and Reserve Bank community.

DANCE development staff identified several descriptive characteristics that they believed would reduce the silo effect. The metadata fields[5] identified the purchased dataset and focused on the following four concepts: description, access rights, contact information, and data location.

As altruistic as DANCE's initial goals may have been, obtaining underlying metadata on each entry proved to be difficult. For example, some units were hesitant to share metadata due to staff workload concerns. DANCE staff populated metadata for as many datasets as possible. However, a thorough metadata entry needed input from users with expertise. Many of the datasets were purchased by the Board's Research and Law Libraries. Thus, efficiencies were gained by having the Research and Law Library staff update and maintain DANCE entries for library purchased resources.

At this point, DANCE contained the majority of vendor purchased datasets, with the remaining spread amongst multiple departments at the Board. Since the datasets were not concentrated within one department, the Library model could not be used. As a result, DANCE staff started working with administrative units within each of the research departments at the Board. Each administrative unit coordinates data purchases for the department, records which datasets are purchased and by whom. Through this, metadata for the Board's remaining purchased datasets were populated in DANCE. Each year following, DANCE staff reconciled entries with each division's administrative staff. While the annual reconciliation was useful in

verifying entries, it was not very timely.

To provide more frequent updates, DANCE staff created a pilot project with one of the administrative units. This project required the data purchasing group to register the dataset with DANCE, which would then auto populate fields and provide information to the respective administrative department. With this enhancement, DANCE contained the most current metadata at all times. This process is planned to be incorporated into the new DANCE structure by implementing similar processes for each division.

Based on consistent usage statistics, DANCE appeared to fit its original purpose to broaden Board use of vendor-purchased datasets. Each user, regardless of department, could view the existence of Board-purchased datasets with ease. Thus, DANCE helped foster cross-departmental collaboration and reduced search costs for purchased data.

DANCE Revisions

In 2006, the research divisions at the Board undertook a new data initiative named the Research Divisions Data Initiative, or RDDI for short. RDDI's purpose was three fold: data cataloging, data storage, and data presentation. Before RDDI could identify and develop better storage and presentation methods, all datasets needed to have standardized documentation created. DANCE had basic documentation for purchased datasets, but the research community wanted enhanced metadata for both purchased and Board created datasets. As a result, a working group on data documentation was convened to identify new metadata fields for DANCE.

The working group used Dublin-core standards as their guide in identifying metadata[6] items to better describe a dataset and thus better facilitate search capabilities. The working group started with over 100 items and whittled them down to 38. Each metadata item was examined for its usefulness in describing a dataset, its potential as a search criterion, and its maintenance difficulty. For example, the Dublin-Core item 'replaces[7]' was removed since the working group thought the marginal cost of maintaining that item would be more than the marginal benefit.

Since DANCE's primary purpose is to document purchased datasets, the administrative units at the Board requested additional items. The administrative units are required to generate frequent reports and memos on dataset costs and corresponding budget justification. To assist with the generation of these internal publications, an additional ten items were added.

Once the new metadata items were established, the working group examined several open source software solutions to display them. Four options were explored: E-prints, Fedora, D-space, and a custom-built solution. Each of the open source solutions presented their own set of installation challenges. For example, D-space runs off of java, however, the Board provides limited java support. Further, the ten additional administrative items were not supported by any of the open source solutions. As a result, the working group concluded that the new DANCE structure would require a custom, Board-built solution.

One of the many benefits of a custom-built solution is flexibility. DANCE was not limited by what metadata items to include or how they were communicated. DANCE further extends this flexibility with added communication methods, security permissions, and variable-level metadata. These enhancements immediately notify users of new datasets or changes to existing datasets.

Each dataset also has a unique set of security requirements. These requirements dictate usage rights, how access is granted, and publication rights. Many times there are paper or electronic access request forms required for dataset use. A custom-built solution allows DANCE developers to display the correct request method for each dataset.

Finally, a custom-built solution enables DANCE to extend beyond dataset metadata to variable metadata. Variable-level metadata gives users a micro-level description of the dataset's contents, thus enabling users to determine a dataset's usefulness for their needs.

New DANCE

The new DANCE is designed not only to accurately describe a dataset, but also around how a user would search/receive metadata related communications, add Board specific documentation, and update metadata items. The search screen will provide multiple ways to search and filter for a dataset. First, a persistent search box will be accessible on all pages, allowing users to search for a different dataset without having to return to the main DANCE page. All the fields will be indexed by the search engine, thus all searches will be free text, unlike library systems which allow users to specify author, title, etc. While the single search box provides a cleaner interface, it provides significant challenges in returning the correct results.

DANCE will incorporate two methods that focus on interaction between the user and the page to meet this challenge. First, the search box will be AJAX-enabled[8] AJAX queries DANCE for datasets that match the typed search without submitting the search. The user can then modify the search string as needed to pull up the desired results. Please see Figure 1 for an example of how each additional letter that is added to a search string reduces the returned results

Second, the user will be able to filter DANCE by keyword or category. While a similar method is used by the

current DANCE iteration, the new version will allow for combining filters. The filtering is designed to help the user drill down to the desired dataset. AJAX search and filtering utilize direct user interaction to yield the desired results. Instead of programmatically trying to determine what the user had in mind, the search methods provide instantaneous feedback so the user may alter search criteria accordingly.

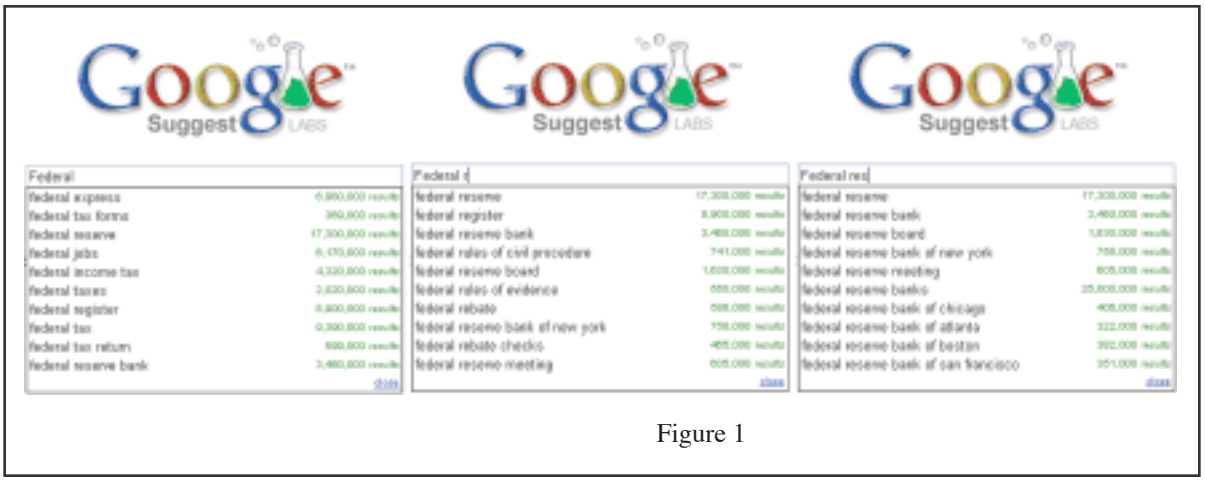

Figure 1

In addition to searching from the main page, DANCE is designed to let a user navigate to different datasets through the result set. Each keyword, category, contact name, etc, is linked to all the results for a particular field. For example, a result for 'industrial production' includes the keyword 'semi-conductor.' A user is able to click on this keyword to see all the datasets that have also been tagged with 'semi-conductor.'

Beyond browsing dataset linkages, users may want to know if there have been any changes to a dataset's metadata or if there are any recently added datasets. The new DANCE will communicate changes and additions through two methods: 1) information boxes on the main search page and 2) a subscription service. Besides the searching and filtering capabilities on the main page, there will be two information boxes with links to the five most recently changed, as well as the five most recently added, datasets. A subscription e-mail service is also available for a user to register for to receive metadata changes or dataset additions.

The new DANCE will also allow for user-generated content through the addition of wiki[9] pages for each dataset. Each dataset's wiki page will be populated initially with vendor-provided documentation and user manuals. Anyone from the Federal Reserve System has the authority to modify a wiki page. With each wiki page, DANCE hopes to capture and share knowledge across the system. In addition to dataset knowledge, DANCE plans to include links to computer code and published papers. Recently, the Board set up a test code repository where anyone can share and critique code. DANCE will link code to the repository designed to manipulate the dataset. To complete the research cycle, DANCE will include links to published papers that use a particular dataset. Currently, each published paper is registered via an online form. DANCE staff will work with the Board's publications department to include a registration field for what datasets were used in the paper.

While the wiki pages allow for anyone to add content, the metadata working group concluded that the core metadata items should only be manipulable by either the data owner or an individual authorized by the data owner. Further, the working group requested that certain core metadata items only be displayed to Board employees[10]. To implement the working group's requests, the new DANCE will use a two-layer security model; one at the application layer and the other at the database layer. The application security will display the authorized fields based upon the user community. The database security will limit metadata editing capabilities to authorized users.

Any user across the Federal Reserve System may request a new entry to DANCE. The new content entry screen will be the same as what is used for the administration dataset registration process. The user's request will be submitted to DANCE staff for review, and added to DANCE after approval. A new feature to the entry form includes fields for the addition of variable-level metadata for the requested dataset.

The new DANCE will accomplish a new level of understanding and collaboration with respect to Board (and potentially Reserve Bank) datasets. The new user interfaces and content distribution methods are designed to easily communicate available data, how to use the data, and published works using the data. While the new DANCE reduces many dataset-specific silos, seamless knowledge transfer is DANCE staff's long term vision.

Data and Knowledge Management Vision
The DANCE working group also requested that DANCE be able to return results from other data search tools, most notably FAME.[11] The group agreed that the most effective search tool would be independent from the metadata or data location, thus fully eliminating Board data silos. In addition to a uniform dataset search, DANCE staff are working with the Board's Research Library to create an enterprise wide search tool.

In this new vision, data, metadata, documents, papers, programs, and library resources will all be indexed and accessible through a universal interface. All result sets will redirect users to the appropriate individual catalogs for more information. Furthermore, a flexible design could include indexing Reserve Bank libraries and resources.

One step toward this integrated vision was taken in spring 2008 with the installation of LibX.[12] LibX is a search-tool that is installed on a browser similar to the quick search Google toolbar. LibX provides one location to search registered repositories; however, each search is repository-specific. For example, a user may want to search 'Stock Prices' in the Board Research Library catalog, DANCE and the E-Journal Portal. LibX requires the user to search each repository separately. While LibX accomplishes the uniform location requirement; a more general enterprise search is still being investigated.

There is already some degree of system coordination. For example, several of the Reserve Bank libraries are searchable through a common gateway. Further, there is a system-wide search. However, if you search for the term 'DANCE' most of the results consist of social dance clubs, dance lessons or hosted events with dancing. While there is still a bit of tweaking to be done, this search could be altered to target research related resources.

Along with simplifying the search process, users must be able to access the datasets. Data storage issues are the next phase of RDDI. An effort is underway to test storing several managed datasets in a relational database server instead of as individual SAS datasets. Relationally stored data may be combined with DANCE to create a data retrieval tool with embedded metadata. The advantage of this setup is that users may build datasets in their preferred format instead of the storage format. The increasing size and number of accessible datasets places further importance on the ability to manipulate data into a research-friendly format.

Conclusions
Economic research is increasingly multi-faceted, and, as a result, research projects are extremely data-intensive. Therefore, metadata documentation and standards have evolved significantly since DANCE's inception. DANCE started as a communication tool identifying the existence of a dataset with minimal descriptive characteristics. DANCE's popularity, along with new data initiatives, led to several metadata enhancement requests to provide deeper searching and administrative capabilities.

Multiple open-source solutions were explored to handle new metadata items and communication requirements. However, technical requirements, as well as additional business requirements, lead to a custom-built solution for the new DANCE. The custom solution enabled new features which focused on the ability to foster collaboration between departments. The new features allow users to seamlessly traverse from dataset metadata, to security request forms, to variable-level metadata without changing applications. These efficiencies will enable researchers and regulators system-wide to focus on data analysis and searching for the data. As a result, it is expected that this focus will translate into even more effective policy making benefiting the entire financial and economic systems..

Appendix 1 – Federal Reserve System Structure

The Federal Reserve System is comprised of the Board of Governors and twelve Reserve Banks. The Federal Open Market Committee (FOMC) is made up of the Board of Governors and presidents of the Reserve Banks. The Board of Governors, in Washington, DC, provides the leadership for the entire system. Both the Reserve Banks and the Board conduct monetary and economic research. The combined Board and Reserve Bank research assists the FOMC in monetary policy decisions.

Appendix 2 – Original DANCE Metadata fields

Metadata Field	Description
Database	The commonly accepted name or title of the database or dataset hyper linked to database specific documentation.
Vendor	The primary vendor name hyper linked to the appropriate website.
Division(s)	The division and co-owning division if applicable.
Form of access	A general description of the network location and access software required.
Status	A Boolean indicating whether the Board still purchases or maintains the database.
Description	An executive summary of the database as well as any Board specific elements.
Keyword(s)	DANCE staff assigned general descriptive words.
Data Contact	The name and phone extension of the primary individual(s) with expertise using the database.
License Contact	The name and phone extension of the individual holding the license agreement.
License Information	One of the Board security classifications or a contact person if express consent is required.
Category	A DANCE staff assigned general data grouping used for search purposes.
Cost	USD cost at the time of purchase.

Appendix 3 –New DANCE Metadata fields

Metadata Field	Dublin – Core Name	Description
Title	Title	The commonly accepted name or title of the database or dataset hyper linked to database specific documentation.
Title Short	Title	Abbreviated name for drop down search
Creator	Creator	The primary dataset creator: could be an internal individual or external company
Publisher	Publisher	The entity that makes the dataset available to the public. May or may not be creator.
Vendor	Vendor	Entity that sold the data to the Board
Data Contact	Data Contact	Primary contact(s) for data questions
Data Requestor	Data Requestor	Individual(s) who requested the dataset
Contributing Section	Contributor	Group(s) with expertise
Data Origination	Source	Original data source; may or may not be the creator or publisher.
Keyword	Subject	Frequently used descriptive terms
Description	Description	Dataset abstract
Date Created	Date	Date the dataset was first created at the FRS.
Date Range Available	Available	The date range the data is available from the vendor.
Geographical Coverage	Coverage	The physical locations the dataset covers.
Type	Type	Abstracted keywords; i.e. micro economics, macro economics.
Data Location		Physical storage location at the FRS.
Output Format	Format	Output file format; example: SAS dataset
Input Format	Medium	Input file format; example: tab-delimited text file
Identifier	Identifier	Auto-generated number uniquely identifying a dataset.
Data Confidentiality	Rights	FRS security classifications assigned to the dataset.
Bibliographic Notation	Bibliographic Citation	How the dataset should be cited in published works.
Related Resources	Relation	Related datasets

Appendix 3 –New DANCE Metadata fields ..(CONT)

License Agreement	License	A scanned PDF of the user license.
License Owner	Rights Holder	The individual(s) responsible for the license
Additional Information	Bulletin Board	The wiki page for the dataset.
Update Method	Accrual Method	Method in which the dataset is updated at the FRS.
Update Schedule	Accrual Periodicity	Frequency that the dataset is updated at the FRS.
Dataset Status	Accrual Policy	An active or inactive flag
Purchasing Division	Division	Division(s) purchasing the dataset.
Purchasing Section	Section	Section(s) purchasing the dataset
Product URL	Product Website URL	Product website
Vendor URL	Vendor Website URL	Vendor website
Vendor Contacts	Vendor Contacts	Technical and sales contacts for the dataset.
Cost	Cost	USD dollar cost
Payment Schedule	Frequency of Payment	Schedule that payments are to be made.
Contract Renewal Date	Contract Renewal Date	The date the contract is to be renewed.
Purchase Justification	Purchase Justification	Budget justification for purchasing the dataset.
Purchase Order	Purchase Order	Number linking the dataset to the procurement purchasing system.
Reason Needed	Need	The economic or regulation reason for purchasing the dataset.
Sole Source	Sole Source	A flag indicating if the dataset is only available from one vendor.
Contract Length	Contract Length	Length of time the contract is valid for.
Record Last Update	Modification History	Log of the time in which a record was modified.
Record Updated By	Catalog Entry Maintained By	Log of individuals who modified a record.

Footnotes

1. The opinions are of the author and not the Federal Reserve Board. Andy Boettcher, Board of Governors of the Federal Reserve System, Washington, D.C. Contact: Andrew.S.Boettcher@frb.gov. This article is based upon a presentation at the IASSIST 2008 conference at Stanford.

2. Market discipline is one of the three pillars of the BASEL II banking accords. Please see http://www.bis.org/publ/bcbsca.htm for details.

3. Fractured data management practices have decreased since data cataloging started and are further reduced with recent data initiatives, described later in the paper.

4. Please see Appendix 1 for brief description of the Federal Reserve System's structure.

5. Please see Appendix 2 – Original DANCE Metadata Fields for a list and descriptions.

6. Please see Appendix 3 – New DANCE Metadata Fields for a metadata list and descriptions.

7..Replaces refines the relationship variable, for example, Dataset B was purchased to replace Dataset A.

8. AJAX stands for asynchronous JavaScript and XML. See http://www.google.com/webhp?complete=1&hl=en for an example.

9. Wikis are software solutions that allow anyone to add or edit content. Please see http://en.wikipedia.org/wiki/Main_Page

10. See Appendix 3 for which user community can view which items.

11. The entire macroeconomic side of research uses FAME for data storage. FAME provides basic metadata for each stored series.

12. LibX is an open-source search tool written by Virginia Tech. www.libx.org