CANNON, S. and PAN, D. 2016. First forays into research data dissemination: a tale from the Kansas City Fed. *IASSIST quarterly* [online], 40(1), pages 35-39. Available from: https://doi.org/10.29173/iq626

First forays into research data dissemination: a tale from the Kansas City Fed.

CANNON, S. and PAN, D.

2016

First Forays into Research Data Dissemination:

A Tale from the Kansas City Fed

by San Cannon[1] and Deng Pan[2]

Abstract

The Federal Reserve System[3] has a longer tradition of doing economic research than disseminating data from economic research. Each of the 12 Reserve Banks and the Board of Governors have research departments that together publish nearly 1,000 working papers and journal articles annually. Unfortunately, researchers have not often made the data from their papers publicly available until recently. A new program at the Federal Reserve Bank of Kansas City aims to correct this imbalance and make such data available for reuse in other research. As a pilot participant in a new dissemination platform, we have educated economists, built metadata specifications, recruited contributors, collaborated with technology and legal staff, and coordinated and built coalitions across multiple functions at our institution and others. This paper outlines the challenges faced and obstacles overcome as we created the infrastructure and workflow and took steps toward making the publication of research data a regular part of the research life cycle.

Keywords

Data dissemination, metadata, usability testing, research data management

Introduction and Background

As researchers are deluged with data, the need for them to share or disseminate their research data widely may not seem pressing—there's plenty to go around. (*The data deluge*, 2010). Researchers now have more choices when compiling data to support their research, and reusing other researchers' data is an important option. A wide range of available data—and advances in analytical capabilities and processing power—means researchers can replicate or build upon a broader variety of research than was possible in the past. Recent mistakes in (and fabrication of) research data have only highlighted the importance of research replicability and the sharing of research data.

The Federal Reserve Bank of Kansas City, like other Reserve Banks and the Board of Governors in Washington D.C., conducts research to support its monetary policy mission, contribute to the safety and soundness of banks, and promote financial stability. The Bank shares its research products with policymakers, other researchers across the System and in academia, and the public. Increasingly, this research requires analyzing vast quantities of data and employing substantial computational resources to address increasingly complex questions.

In Kansas City, more than two dozen researchers and research associates produce about 50 research products (journal articles, working papers, etc.) each year. They are part of a larger community of more than 750 researchers across the System who produce nearly 1,000 such works each year. As the need to acquire data inputs to this research has increased, so has the pressure on support staff and the budget. To help alleviate some of the pressure, the 12 Reserve Banks and the Board have been formally collaborating to acquire source data as research inputs.

While the System has established a set of services to bring in data, there are no such services for pushing out data once the research is complete. The Federal Reserve currently has no coordinated approach to preserving and disseminating research data across the System, and differences in strategies and resources have precluded a consolidated approach. Other academic domains face similar challenges (Borgman, 2012).

Challenges and Opportunities

Setting up a repository or an archive, or defining a workflow to support data preservation or future dissemination, are not just technology decisions. For many disciplines, these activities require a fundamental change in researchers' perceptions of the research process. For most researchers, the ultimate goal is publication; Fed economists are no exception. All data-related work is simply in support of that goal: 'Time and money spent on documenting data for use by others are resources not spent in data collection, analysis, equipment, publication fees, conference travel, writing papers and proposals, or other research necessities' (Borgman, 2012). Effecting change in such circumstances is difficult but not impossible, and research funders may lead the charge. For example, some funding agencies now require publicly funded researchers to make their underlying data available to the public. Although these requirements do not affect Federal Reserve researchers, they will undoubtedly help change the culture of empirical research as a whole (Arzberger et al., 2004).

Even without funding requirements, a few banks in the Federal Reserve System have begun considering how to disseminate their research data sets. The Federal Reserve Board, for example, publishes data for select working papers along with the papers on their website[4]. The data are being disseminated, but researchers must know with which paper they are associated and must then go to that page. The Federal Reserve Bank of New York takes a similar approach but is also compiling a separate page for such data[5]. The Federal Reserve Bank of Kansas City, however, does not currently disseminate research data sets on its public website. For many Reserve Banks, the major means by which research data is disseminated is individual requests to the author. This process, though widespread even in the academic community, is taxing on the author and does not encourage broader reuse.

A coordinated approach to data acquisition in the Federal Reserve System began in late 2011 and was aided by the creation of a data librarian role in each of the Reserve Banks. The Federal Reserve Bank of Kansas City took research support one step further by creating the Center for the Advancement of Research and Data in Economics (CADRE)[6] in early 2015. CADRE's mission is to support, enhance, and advance data or computationally intensive research in economics. Bank leadership identified research data preservation and dissemination as important support functions for CADRE. As these functions were being designed, the Kansas City Fed was offered, and accepted, an opportunity to join a pilot program to help provide use cases and product suggestions for a new research data dissemination platform being developed by a nonprofit academic partnership outside the Federal Reserve System.

Becoming a Pilot Participant

The new publication platform is designed for researchers to submit data directly for dissemination. The workflow in the platform allows for two stages: an initial submission by the researcher and a curation step to verify, edit, update, or clarify the contents of the initial submission before publication. To participate, CADRE needed to specify the details of each stage in the workflow and build use cases for our research community. To do so we considered the following questions:

- What kind of collections? In the dissemination platform, a collection is a group of data sets for which similar policies and access controls can be set. Because metadata and access rights are controlled at this level, we created a public data collection and a restricted access data collection based on our evaluation of the data sets that had been used by researchers in the Federal Reserve System. We needed to be able to store data to which access was limited and test if the access controls worked sufficiently to meet our information security requirements. We also wanted to test the submission workflow and applicability of common metadata across the collections.
- What kind and size of data files? Although the platform was built to accommodate very large files stored in a central location, the typical data files in the pilot range from 1MB to 5GB, and are stored on the researcher's desktop or on the Kansas City Fed's high-performance computing cluster.
- What kind of workflows? In the initial phase of the submission workflow, submitters fill in required metadata fields to describe the data sets and then assemble data files. In the subsequent curation phase, curators review and possibly modify the metadata or files before approving or rejecting the submission.

In addition, we had some practical questions about how the pilot itself should proceed.

- Who would be the testers? We expected four to six test users, mostly economists and their research associates at the Kansas City Fed, to start this pilot and provide initial feedback. We also hoped to expand the test-user base to include users in a few other Reserve Banks as well as their co-authors at academic institutions.
- Who would be the curator? While CADRE had plans to hire a Data Curator, the position had not yet been filled. The curator role was thus temporarily filled by two staff (specifically, the authors).

Metadata Creation

The most critical decisions for this pilot involved choosing which metadata fields the data set submissions should capture. Opinions vary on how much information is sufficient to adequately describe any item. We investigated several existing specifications to evaluate what others view as necessary information for finding or discovering data.

First, we looked at the requirements for depositing a data set at the University of Michigan's Inter-university Consortium of Political and Social Research (ICPSR), which maintains a data archive for social science research data[7]. Although our data-dissemination pilot is not

meant as an archive, many of the issues for data discovery are the same. The submission process for ICPSR doesn't explicitly require any fields, but we believe Title, Principle Investigator, and Description are the minimum metadata fields that are practicable. In addition, ICPSR requests another dozen or so categories of metadata ranging from mode of collection to geographic coverage. Because the ICPSR archive handles a large number of data sets that are primary data collections, many fields, such as response rate and weights, rarely apply to data used in Federal Reserve research projects.

Next, we reviewed one metadata schema specifically designed to help users discover data. For Federal agencies that provide data, the executive order OMB 13-13 specifies a particular metadata schema to catalog data assets[8]. The metadata elements defined for that order are published as part of Project Open Data[9] comprise 12 required fields, some common to other specifications (for example, title, description, keyword) and others specific to government agencies (Bureau Code, Program Code). In addition, six fields are required by the schema if applicable, including license, rights, and spatial and temporal metadata.

Finally, we examined the metadata requirements that an internal workgroup had developed to catalog data assets across the Federal Reserve System. This unpublished specification lists more than 30 metadata elements including 14 mandatory elements. Some items are common to other specifications (such as name or description), whereas others describe access restrictions (for example, security classifications).

After carefully considering the user burden for metadata entry, we decided on the following metadata elements.

Required	Optional
Data set name	Additional information
Contact author name	Key words
Contact author information	Journal of Economic Literature Classification
Description	Geography
Update information	Unit of observation
Category	Date(s)
Frequency	Documentation
Access restriction	
Security classification	

We also added three metadata fields to the curation workflow: file type(s), file size(s), and article information. The file specification information is important both for technical staff managing the file space and for users who initiate a download. Incorporating these into the curation workflow rather than the submission process reduces burden on the depositor and allows for possible changes to file types and sizes (through compression, for example) before the data are published.

Infrastructure Choices and Challenges

Once the interface and workflow were developed to our specifications, we performed initial testing before involving our users. We faced early infrastructure challenges; specifically, getting the dissemination platform to work well with our environment. The dissemination platform is meant to interact with our existing file system and storage infrastructure. We had some difficulty getting the file system and publication platform to interact smoothly. Because we need to allow access by outside users, the file storage and platform live in an external zone of the Kansas City Fed's intranet. This positioning made it easier for external users to get to the files, but made it more challenging for internal users to load the files from their desktop to the endpoint. The challenge was not insurmountable, but it did frustrate us as we worked through a typical user experience.

Another major infrastructure difficulty involved authentication and identity management. As part of a pilot project, we needed to have accounts on the platform infrastructure. These identities were then used to manage access to the collections through group definitions. For our initial work, adding individual users to the appropriate roles and access groups was fairly straightforward. However, in planning for a more robust long-term implementation, we do not want to maintain identities and security groups separate from existing information security infrastructure.

Usability Testing

The goal of this pilot is to evaluate whether this data dissemination platform meets our expectations from both the technical and user perspectives. How useable Bank and System staff will find any tool is one of our primary concerns. While we were not able to engage in

any formal usability testing, we did want to get informal feedback from potential users. We worked with three volunteers to address the following questions.

- How long do users need to complete the entire submission workflow and do they consider the process burdensome?
- Do the metadata fields defined in the pilot describe the data sufficiently?
- Are users willing to provide supporting documents and further descriptions, such as a data dictionary, in the submission process?
- How effective is the user interface? Do users suggest any improvements?
- What is the role of the data curator and how could this individual help improve the efficiency of the workflow?

Each user completed the test independently to diminish peer influence. During the test, we instructed each participant to walk through the submission workflow: logging in, entering the necessary metadata, and then assembling and submitting the data file. Once the test was completed, the participant was asked to provide feedback on his or her overall impression of the tool and the effectiveness of the workflow, as well as suggestions on how to improve the user interface.

Users found some steps in the workflow challenging at first. They received registration emails for access to the platform that they weren't sure how to handle and for which we had failed to prepare them. Even once they understood the registration process, they were somewhat stymied by a technical difficulty peculiar to this pilot: the submission process relied on two separate websites for different parts of the workflow. During the pilot, the two sites were not well integrated, resulting in some confusion as users navigated between two similar looking, but disconnected, web pages. We do not expect this problem when the platform is commercially available.

Once the users understood where to start, we observed them while they completed the two parts of the submission workflow: metadata entry and data file assembly. Metadata entry required the users to fill out all of the mandatory fields and presented optional fields as well. By limiting the number of required entries, we hoped to improve efficiency and reduce the cost to researchers for publishing their data. None of the users we observed seemed to notice the distinction between required and optional, so they simply filled out all of the blank fields.

The next step, data file assembly, required users to upload data from their computers to the Kansas City Fed staging area for the platform. As previously mentioned, users encountered some technical difficulties with the connection to the storage system, and only two of the three testers were able to successfully construct and upload a data file.

User Feedback

Overall, the users reported that the amount of time spent on submission was not burdensome. They also reported that a majority of the metadata fields were easy to fill out. However, the users offered a few suggestions regarding metadata and user interface at the end of the test:

- Enhance the metadata selection interface
  Users were asked to select terms from a controlled vocabulary specific to economics and finance. To encourage consistency across the Federal Reserve System, we used a list maintained by colleagues at the Federal Reserve Board of Governors to describe their research publications. The list was a flattened hierarchy of more than 70 lines and was difficult to navigate in a drop-down menu. Users suggested retaining the hierarchical structure but splitting the long list into multiple drop-downs to improve navigation and readability.

- Add detail to some fields
  The Journal of Economic Literature (JEL) maintains an alphanumeric, hierarchical classification system that is the standard for classifying scholarly literature in the field of economics[10]. The interface for submission used the text description for the economic field without including the 2-3 digit identifier. Because economists are so familiar with the JEL scheme, users suggested adding the code designation next to the description.

- Clarify metadata fields
  The users found a few metadata fields ambiguous and suggested clarifications and examples to improve the workflow. For instance, we named one metadata field 'Data Restriction Audience' to contain information on time restrictions (when can the data be shared) and access restrictions (with whom can the data be shared). Users weren't sure what the name implied, and they certainly did not understand our intended usage.

  The field labeled 'Documentation' also confused users. We expected users to provide supporting documentation such as a data dictionary. One economist expressed willingness to do so but was unsure how much detail was required. He also pointed out that as the definition of certain data variables has changed over the years, providing information on how the data was constructed and elaborating on the difference would add high value to data dissemination.

- Allow options for restricted data
  Many of the researchers in the Federal Reserve System are assigned to work with restricted data that can only be shared with certain audiences. The user who volunteered to test the platform using restricted data noted that some of the data could still be made available after sensitive information was extracted. Creating versions of the data that can be shared more broadly will definitely

require more work on the researchers' side, but will eventually be beneficial to the research community within the Federal Reserve System as well as to the public.

- Ensure persistent identification
  All three test users expressed concerns with the permanence of the URL for their research output. As the purpose of data dissemination is to make data available, the economists wanted to know how we would ensure consistent access to the data once published. CADRE staff were already working on a program to assign digital object identifiers (DOIs) to research output, which would ensure current location information for files and provide identifiers to published datasets.

Curation Testing

In the curation workflow, the curator reviews the submitted metadata and the data file, adds additional information as needed, and approves or rejects the submission. Though we had early success in curating entries in the pilot, we were not able to curate the data the three users submitted during this phase of testing. There appeared to be some technical issue that neither the platform developers nor the Kansas City staff could identify or resolve. After a seemingly unrelated patch application, the problem disappeared. All curation attempts thereafter were successful.

Overall, the curation workflow functioned as anticipated and worked well. We are considering changing the workflow so that the curator assembles the data file instead of the data submitter but have not yet tested this possibility.

Next Steps

To accomplish the objectives set at the beginning of the pilot, we need to take the following steps:

- Repeat the usability testing with the three economists to ensure all of them are able to submit metadata and upload data files from their computers. We will modify the curation testing to determine whether the curator is able to assemble data files for the data submitter and to ensure the submitted metadata and data are successfully curated.
- Finalize the dissemination workflow to include DOI assignments which were created locally and temporarily during the pilot. We will investigate how the publication platform integrates with our DOI registration service.
- Work with technical staff to verify the security settings of the platform. Ideally, we would implement a data dissemination platform for both public data and restricted access data. One important aspect of this pilot is evaluating the access settings of the platform to ensure they meet our security requirements.
- Expand this pilot to other interested participants such as researchers in other Federal Reserve Banks and the Federal Reserve Board of Governors. We anticipate involving more potential users and collecting feedback from them, particularly on their experiences with metadata entry, data submission, and identity authentication.

References

Arzberger, P. et al. 2004, 'Promoting access to public research data for scientific, economic, and social development', Data Science Journal, vol. 3, no. 29, pp. 135–152. Available from:
<https://www.jstage.jst.go.jp/article/dsj/3/0/3_0_135/_pdf>. [24 June 2015].

Borgman, C. 2012, 'The conundrum of sharing research data', Journal of the Association for Information Science and Technology, vol. 63, no. 6, pp. 1059–1078. Available from: <http://onlinelibrary.wiley.com/enhanced/doi/10.1002/asi.22634>. [24 June 2015].

The data deluge. 2010. Economist. Available from: <http://www.economist.com/node/15579717>. [24 June 2015].

Notes

1, San Cannon, Federal Reserve Bank of Kansas City. Email: Sandra.Cannon@kc.frb.org.

2, Deng Pan, Federal Reserve Bank of Chicago. Email: Deng.Pan@chi.frb.org.

3. http://federalreserveonline.org/

4. http://www.federalreserve.gov/pubs/feds/2005/200533/200533abs.html

5. http://www.newyorkfed.org/research/staff_reports/sr493.html

6. https://www.kansascityfed.org/research/cadre/

7. http://www.icpsr.umich.edu/icpsrweb/deposit/

8. https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf

9. https://project-open-data.cio.gov/v1.1/schema/

10. https://www.aeaweb.org/econlit/jelCodes.php