Plan Overview

A Data Management Plan created using DMPTool

Title: NSF CAREER Grant

Creator: Adrien Matray

Affiliation: Princeton University (princeton.edu)

Funder: National Science Foundation (nsf.gov)

Funding opportunity number: 000833860

Template: NSF-SBE: Social, Behavioral, Economic Sciences

Project abstract:

This project will study how financial markets affect economic development by focusing on the role of the misallocation of capital across individual, firms, sectors and places, as well as the types of government interventions in financial markets that can successfully ensure capital is well allocated. The first research objective is to study the U.S. in the 19th century, and the second is to study Brazil in the modern day. The study of the U.S. will empirically investigate how financial development affects the misallocation of capital across: farms in agriculture (Project 1), firms in manufacturing (Project 2), and individuals investing in human capital (Project 3). The study of Brazil will empirically investigate two major government interventions in the banking sector. One aimed at promoting financial inclusion (Project 4) and one aimed at sustaining firms' exports during the financial crisis of 2008

Start date: 02-01-2021

Last modified: 08-07-2020

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customize it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

NSF CAREER Grant

Roles and responsibilities

The DMP should outline the rights and obligations of all parties as to their roles and responsibilities in the management and retention of research data. It should also consider changes to roles and responsibilities that will occur should a principal investigator or co-PI leave the institution or project. Any costs should be explained in the Budget Justification pages.

I will personally manage all the data components of this project.

Expected data

The DMP should describe the types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project. It should then describe the expected types of data to be retained.

Research Objective 1: The U.S. 1850-1960

This research objective will collect a large amount of archival - publicly available data:

Annual bank-level balance sheets. These data will come from two sources: (i) National Bank Balance Sheets: Comptroller of the Currency - 1865-1934; (ii) State Bank Balance Sheets: various State Banking Department Records- 1863-1934. Variables of interest available in these balance sheets include the different components of assets (total loans, short-term investments like US bonds, cash, transfers from other banks, etc) and liabilities (capital, deposits, debt, transfers to other banks, etc). Each bank will be matched to its county and thereby linked to the US census data. While some of these data have been collected for sub-periods, these data have not been collected in a systematic way for the whole period 1850--1940, are not publicly available and have not been matched to the manufacturing sector in the proposed way. This will create the most comprehensive panel linking local US financial conditions to economic outcomes, historically.

Agriculture data:

- Comparative advantage: expected output at the crop-county level will come from the FAO GAEZ database.
- True agricultural output data per county will come from the Census of Agriculture, supplemented with annual data on county level crop data found in Statistical Reports of multiple State Boards of Agriculture.

Aggregate county-industry: data will come from two sources:

- Census of Manufacturing (1850, 1900-1930, 1954, 1958, 1972), using the census data collected by Rick Hornbeck and Martin Rotemberg for 1860--1880, and collecting the Annual survey of manufactures (issued biennally for the period 1949/1950--1980/1981). It is possible to construct a panel at the county-by-industry-by-survey year level that contains the total value of products (or ``total output''), the total value of materials used (fuel and other ``inputs'') and the total amount paid in wages during the year (``labor inputs'').
  Firm level: similar information on the output, input and capital can be retrieved at the firm-year level after 1900 from Moody's Industrial Manuals, which provide annual firm balance sheet for a large number of manufacturing firms. The manual was for investors and collected and reported balance sheet information of all companies that had issued a bond or a stock at some point.

Firm financial strength. Firm credit score and net worth from Dunn and Bradstreet historical manual. Collecting the universe of firms will probably be unfeasible. I will therefore collect random sample for all the main counties existing since 1850 at every ten year interval as well as for the year before each major financial crisis (1893, 1907, 1920, 1928).

De-anonymized full count census for 1850-1940: the de-anonymized version is available on the NBER server.

Patent data: yearly level of all patents filed at the USPTO for the period 1850-1940. The name of inventors will be extracted from the patent text and will be merged with the Census full count using names. The final match will result in a database linking each patent to individual using the ``hist-id'' identifier, publicly available on IPUMS website, as the de-anonymized version cannot leave the NBER server.

Research Objective 2: Brazil in the modern days

City-level financial services & profitability. Individual bank balance sheets for branches at the city-year level are publicly available and contain information about city-level loans and deposits; individual branch profitability and level of bank risk both for government-owned banks and private banks.

These data once cleaned can be freely shared.

Individual outcomes. I will use the matched employer-employee data with panel information about employees containing occupation, wages, number of hours worked, sociodemographic variables such as sex, age, education. The data are available since 1988 and are accessible via a secure server at Princeton University.

While it is impossible to share publicly individual level data, the projects on Brazil will results in the production of many statistics at the city-year level (e.g. average wage, income distribution, share of unskilled and skilled, industrial specialization, etc.) that can be shared publicly.

Firm-level outcomes:

- Universe of exports and imports data at the firm-product-destination year level are available through a remote server via Marc Muendler at USC. The data contains a unique firm identifier that allow to match custom data at the firm level with the matched employer-employee data, the credit registry and firm balance sheets. Again, these data cannot be shared publicly but aggregate version of the data at the city-industry-year level will be made available.
- Firm balance sheets come from *Pesquisa Industrial Anual da Empresa* (PIA), which is based on annual surveys filled out by firms in the manufacturing and mining sector. These data contain information on operational and non-operational costs, revenues, assets, and investments that allows to estimate firm productivity. The data can be accessed at the Central Bank and the Statistical Office via research agreements.
- Credit registry linking firms and banks at the loan level can be accessed at the Central Bank and the Statistical Office via research agreements.

Period of data retention

SBE is committed to timely and rapid data distribution. However, it recognizes that types of data can vary widely and that acceptable norms also vary by scientific discipline. It is strongly committed, however, to the underlying principle of timely access, and applicants should address how this will be met in their DMP statement.

All publicly available data will be made available for the timely and rapid distribution as they are collected and cleaned.

Data format and dissemination

The DMP should describe data formats, media, and dissemination approaches that will be used to make data and metadata available to others. Policies for public access and sharing should be described, including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements. Research centers and major partnerships with industry or other user communities must also address how data are to be shared and managed with partners, center members, and other major stakeholders.

All historical data need to be digitized at the exception of the Census full count and the Census of Agriculture.

All the OCR archives files will be preserved and made available upon request and will correspond to jpg photographs and scanned that will be saved in pdf format.

All the data to be exploited with statistical software will be saved in excel spreadsheet format, comma delimited text files and in Stata data set format and made publicly available on my website. This will result in a long panel at the county level containing all the variables that will be collected and cleaned for the realization of the different projects (e.g. average bank leverage, deposit, output by industry, output of each crop, number of patents, etc.). These data will be available for scholarly research and other general public use and should not be subject to any copyright protection to the best of our knowledge. They will be available from a specific website that I will design to host and distribute all the material associated with the NSF CAREER Grant.

Micro data on Brazil cannot be shared at the exception of the banks' balance sheet. However, as described above, several aggregate variables will be produced at the city x year level describing the local labor market (average wage, number of workers by industries, wage distribution, share of skilled / unskilled, skill premium, etc.) as well as export and imports and will be made publicly available on my website. In addition, all the cleaning files to organize the raw data and prepare them for statistical analysis (which requires a large investment) can be shared and will be made publicly available.

I will provide a year-by-year listing of the availability of each specific data items. The data will be managed and stored in a spreadsheet format with each column heading providing indicative variable names. Further description of the data will be available in text files describing the source, definition and value labels for all available variables. In addition to providing spreadsheet formats of the data we can make the data available in tab delimited .txt files and in stata format files.

Data storage and preservation of access

The DMP should describe physical and cyber resources and facilities that will be used for the effective preservation and storage of research data. These can include third party facilities and repositories.

All publicly available data will be stored on my personal secured server at Princeton University. They will also be stored on the website that will be designed specifically for these data.

In addition, the will be put on ICPSR.

Census full count de-anonymized are stored on the NBER server.

Brazilian matched employer-employee data will be stored on Princeton secured server.

Brazilian export – import data are stored on a server at USC.

Firm balance sheets and credit registry data are stored at the Brazilian Central Bank.

Because the data collected will be stored in multiple formats (.txt, .xlsx, .dta) maintenance of the data will be unnecessary. ASCII formats will likely be the most usable over the long-run and will require little maintenance.

Additional possible data management requirements

More stringent data management requirements may be specified in particular NSF solicitations or result from local policies and best practices at the PI's home institution. Additional requirements will be specified in the program solicitation and award conditions. Principal Investigators to be supported by such programs must discuss how they will meet these additional requirements in their Data Management Plans.