SOCIOLOGY / DEMOGRAPHICS – DATA MANAGEMENT PLAN

DESCRIPTION

This is a fictional exemplar data management plan developed in May 2020 by Alex Guindon (Geospatial and Data Services Librarian, Concordia University) for educational and guidance purposes. It is mostly based on information taken from a data curation profile where researchers describe how they "use data from the US Census and other federal and state government sources and process, analyze, and distribute the data on the project website in a way that makes the data more accessible and easier to use" (Jenkins, 2012).

Jenkins, K. (2012). Sociology / Demographics - Cornell University. *Data Curation Profiles Directory*, *4*(6). https://doi.org/10.5703/1288284315013

DATA COLLECTION

WHAT TYPES OF DATA WILL YOU COLLECT, CREATE, LINK TO, ACQUIRE AND/OR RECORD?
The collected data is tabular data from the U.S. Census and Bureau of Labor Statistics and the New York State Department of Health. Based on that data, we will then produce additional data which will take the form of a relational database. The final product will be a dynamic website that will include maps, tables and charts.

WHAT FILE FORMATS WILL YOUR DATA BE COLLECTED IN? WILL THESE FORMATS ALLOW FOR DATA RE-USE, SHARING AND LONG-TERM ACCESS TO THE DATA?
The original data tables are publicly available in xls,.csv, .html, .shp and fixed-width text files. The data is then treated (augmented and aggregated) and stored in a MS SQL format. The data output is currently shared as html ((with embedded, tables, charts, trendlines and maps) and in .xls format. The researchers will try to make open formats (.xlsx or .csv) available for the public table.

WHAT CONVENTIONS AND PROCEDURES WILL YOU USE TO STRUCTURE, NAME AND VERSION-CONTROL YOUR FILES TO HELP YOU AND OTHERS BETTER UNDERSTAND HOW YOUR DATA ARE ORGANIZED?
The original data files are already organized and named according to the conventions used by the Census and Bureau of Labor Statistics. Additional data files that will be created (estimates and projections) are named based on these conventions with additional information (type of analyses and versions) appended to the file name. Examples: NYC_Census_2000_estimate_v2; NYC_Census_2000_project_v3.

DOCUMENTATION AND METADATA

WHAT DOCUMENTATION WILL BE NEEDED FOR THE DATA TO BE READ AND INTERPRETED CORRECTLY IN THE FUTURE?
The original data is fully document on the Census and Bureau of Labor Statistics websites.
The metadata describing the variables and the source of data tables is stored in specific "metatables" in the MS SQL database. These are regularly exported as .txt files. In addition, the complete methodology employed to create the estimates and projections is documented and available in .docx and PDF formats. SPSS Syntax files are also provided for the estimates and projections.

HOW WILL YOU MAKE SURE THAT DOCUMENTATION IS CREATED OR CAPTURED CONSISTENTLY THROUGHOUT YOUR PROJECT?
Whenever a new methodology is created (to produce an estimate or a projection, or to create a map using ArcGIS) it is immediately documented by the researcher responsible for these files. A research assistant is in charge for checking that the new methodology is added to the main methodology guide and that all original data produced by the team is properly documented.

IF YOU ARE USING A METADATA STANDARD AND/OR TOOLS TO DOCUMENT AND DESCRIBE YOUR DATA, PLEASE LIST HERE.
All new files created are coded using the Data Documentation Initiative (DDI) metadata standard. To help with the DDI coding. the research team will use the Colectica for Excel software.

STORAGE AND BACKUP

WHAT ARE THE ANTICIPATED STORAGE REQUIREMENTS FOR YOUR PROJECT, IN TERMS OF STORAGE SPACE (IN MEGABYTES, GIGABYTES, TERABYTES, ETC.) AND THE LENGTH OF TIME YOU WILL BE STORING IT?
This is an ongoing project that collects and produces data annually. If we plan for the 10 next years of data, at a yearly rate of 500 KB for the tabular data and approximately 100 MB for shapefiles, we can estimate a total of 5 MB of tabular data and 1 GB of geospatial data.
We think that the data may remain useful for 50 years, so storage should be available for that duration.

HOW AND WHERE WILL YOUR DATA BE STORED AND BACKED UP DURING YOUR RESEARCH PROJECT?
We plan to store our working files on Cornell Box.
This institutional service backs up the files automatically and does automatic versioning while keeping the 100 most recent versions.

HOW WILL THE RESEARCH TEAM AND OTHER COLLABORATORS ACCESS, MODIFY, AND CONTRIBUTE DATA THROUGHOUT THE PROJECT?
Work on data files is either done directly on the shared files in Cornell Box or, if not possible, are copied back to Cornell Box daily. Both data storage and data transmission are encrypted. Team members, either from Cornell or from other institutions will be given password protected access. The data in Box is automatically audited to show who has acted on and viewed our files.

PRESERVATION

WHERE WILL YOU DEPOSIT YOUR DATA FOR LONG-TERM PRESERVATION AND ACCESS AT THE END OF YOUR RESEARCH PROJECT?
For the public data (the vast majority of files), we will consult with the Cornell Research Data Management Service Group to determine which repository platform would best accommodate long-term preservation and access through Google Charts and Google Maps APIs for dynamic visualization of data on our website. Possibilities include Cornell Box and Google Drive (Cornell G Suite) .
The few restricted data files (preliminary, non-public census data) will be stored with the Cornell Restricted Access Data Center (CRADC).

INDICATE HOW YOU WILL ENSURE YOUR DATA IS PRESERVATION READY. CONSIDER PRESERVATION-FRIENDLY FILE FORMATS, ENSURING FILE INTEGRITY, ANONYMIZATION AND DE-IDENTIFICATION, INCLUSION OF SUPPORTING DOCUMENTATION.
Except for the preliminary census files that are restricted and will not be made available publicly, all other datasets will be made available in open formats such as .csv or .xlsx. The source data (from the Census and Bureau of Labor statistics) is already aggregated and has been checked to prevent re-identification of respondents. Therefore, no anonymization or de-identification will be necessary.
The methodology used to produce estimates and projections (in the form of docx or txt documents), along with SPSS syntax files, will also be available from the website.

SHARING AND REUSE

WHAT DATA WILL YOU BE SHARING AND IN WHAT FORM? (E.G. RAW, PROCESSED, ANALYZED, FINAL).
All the data we produce (charts, aggregation, estimates) will be made publicly available on our website. It will also be available for download in .csv and .xlsx formats for the tabular data and in shapefile (.shp) format for the maps. All of these outputs constitute analyzed and aggregate data. We hold no raw data (census microdata for instance) because all that we collect is already aggregated by the Census Bureau or the Bureau of Labor Statistics.
The only data that will be restricted is the census preliminary estimates that will not be shared at all outside of the research team.

HAVE YOU CONSIDERED WHAT TYPE OF END-USER LICENSE TO INCLUDE WITH YOUR DATA?
All the data will be available publicly with a CC0 license. We would appreciate that other researchers properly cite our data, but this is not an absolute requirement.

WHAT STEPS WILL BE TAKEN TO HELP THE RESEARCH COMMUNITY KNOW THAT YOUR DATA EXISTS?
Our website is indexed by all the major browsers. We will provide ready-citations for our files based on the recommendations from the Digital Curation Centre. These citations should be used for publications based on our data.

RESPONSIBILITIES AND RESOURCES

IDENTIFY WHO WILL BE RESPONSIBLE FOR MANAGING THIS PROJECT'S DATA DURING AND AFTER THE PROJECT AND THE MAJOR DATA MANAGEMENT TASKS FOR WHICH THEY WILL BE RESPONSIBLE.
Every researcher in the project team will be responsible for describing the original data they create (using the DDI schema and appropriate methodological documentation). A research assistant (RA) will check, on a weekly basis that all files/dataset have been properly described and documented. This will necessitate a basic training of the RA and the PIs in the use of DDI and the Colectica for Excel software.

HOW WILL RESPONSIBILITIES FOR MANAGING DATA ACTIVITIES BE HANDLED IF SUBSTANTIVE CHANGES HAPPEN IN THE PERSONNEL OVERSEEING THE PROJECT'S DATA, INCLUDING A CHANGE OF PRINCIPAL INVESTIGATOR?
As the data will always reside on Cornell's Box and will be continuously documented, there should be no serious consequence if a member of the research team leaves the project. But if the lead researcher leaves, then one of the other PIs will take over the responsibility for the data. If the RA leaves before the completion of the project, a new RA will be trained to perform the metadata quality check.

WHAT RESOURCES WILL YOU REQUIRE TO IMPLEMENT YOUR DATA MANAGEMENT PLAN? WHAT DO YOU ESTIMATE THE OVERALL COST FOR DATA MANAGEMENT TO BE?
These costs are planned for the 4 years of the New York's Empire State Development (ESD) funding. The total size of the dataset I'ss modest and the storage and long-term preservation costs are covered by Cornell's IT service. We estimate the salary of the RA at $ 400/week (20 hours a week at an hourly rate of $ 20) for 30 weeks a year, for a total of $48,000 for 4 years. Other costs may include occasional online training on DDI and/or Colectica. We estimate that these will not be over $ 500 per year on average for a total of $2,000 over 4 years.

ETHICS AND LEGAL COMPLIANCE

IF YOUR RESEARCH PROJECT INCLUDES SENSITIVE DATA, HOW WILL YOU ENSURE THAT IT IS SECURELY MANAGED AND ACCESSIBLE ONLY TO APPROVED MEMBERS OF THE PROJECT?
The only sensitive data (the Census preliminary estimates) will be kept on the Cornell Restricted Access Data Center (CRADC) and only accessible to the PIs. Access is granted via Cornell username and password. The data on CRADC is encrypted.

IF APPLICABLE, WHAT STRATEGIES WILL YOU UNDERTAKE TO ADDRESS SECONDARY USES OF SENSITIVE DATA?
There will be no secondary use of sensitive data.

HOW WILL YOU MANAGE LEGAL, ETHICAL, AND INTELLECTUAL PROPERTY ISSUES?
All of the collected data remains the property of the Census Bureau and the Bureau of Labor Statistics. The estimates, projections and aggregations created by the research team will be made available under a CC0 license and will thus be part of the public domain.