Collecting Data (Assignment 2)

Horizon Europe
Data Management Plan

22 November 2023

Data Management Plan created in Data Stewardship Wizard

History of changes

There are no named versions.

Contributors

The following contributors are related to the project of this DMP:

- Jennifer Dijkstra
  j.dijkstra.56@student.rug.nl
  Roles: Contact Person, Data Curator, Data Manager
  Affiliation:

  University of Groningen

Projects

We will be working on the following projects and for those are the data and work described in this DMP.

Irish-American Census

Acronym
CD-IAC

Start date
N/A

End date
N/A

Funding
Did not apply for any funding yet.

This projects aims to visualise the population of Irish-American persons in the United States of America.

1. Data Summary

Re-used datasets

We have found the following reference datasets that we have considered for re-use:

- **Individuals of Irish ancestry (American Community Survey 1-Year Summary Files)** (https://data.diversitydatakids.org/dataset/04006_5_c-individuals-of-irish-ancestry--count-) ✓

  Owner of this dataset: DDK (diversitydatakids.org) is a project of the Institute for Child, Youth and Family Policy at Brandeis University, a not-for-profit Massachusetts corporation ("Brandeis") and owned by Brandeis. info@diversitydatakids.org / 415 South Street, Waltham, MA 02254.

  The dataset can be used in the provided format without any conversion needed.

  We will use version "Version 1." of this dataset. If a new version becomes available during the project, new analyses will be done with the new version.

  We will keep a copy of the dataset and make it available with our results for the reproducibility.

  We will use the dataset as follows: To visualise the population of Irish-American persons in the United States of America on a national, regional, state, and city level.

There is no need to harmonize different sources of existing data in our case.

Data formats and types

We will be using the following data formats and types:

- **Comma-separated Values** (CSV)

  A comma-separated values (CSV) file is a delimited text file that uses a comma to separate values. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format. A CSV file typically stores tabular data (numbers and text) in plain text, in which case each line will have the same number of fields.

  It is a standardized format. This is a suitable format for long-term archiving. We will have only a small amount of data stored in this format.

2. FAIR Data

2.1. Making data findable, including provisions for metadata

There are no 'Minimal Metadata About ...' (MIA...) standards for our experiments. However, we have a good idea of what metadata is needed to make it possible for others to read and interpret our data in the future.

We will use an electronic lab notebook to make sure that there is good provenance of the data analysis.

We made a SOP (Standard Operating Procedure) for file naming. Folders should be sorted by their code first. for example: "010_nation", "040_states", etc. files should be ordered by date created. for example: "2023-22-11_nation.csv". We will be keeping the relationships between data clear in the file names.

2.2. Making data accessible

We will be working with the philosophy *as open as possible* for our data.

All of our data can become completely open over time.

Limited embargo will not be used as all data will be opened immediately.

Metadata will be openly available including instructions how to get access to the data. Metadata will available in a form that can be harvested and indexed (managed by the used repository / repositories).

Our data is legally not copyrightable, there is no legal owner.

For the reference and non-reference data sets that we reuse, conditions are as follows:

- **Individuals of Irish ancestry (American Community Survey 1-Year Summary Files)** – available under specific restrictions, which we will follow in our project:

  Brandeis grants an end-user, a revocable, non-exclusive, non-transferable, non-sublicensable limited right and license to access and use the Original Content solely for their own personal research use. They are are authorized to view, play, print, and download Original Content found on the Site for personal, informational, research or non-commercial purposes only, provided that you keep all copyright or other proprietary notices intact. (BRANDEIS copyright notice shall thereon be included with any publication using the Original Content. Such notice shall be affixed so as to give reasonable notice of Brandeis's claim of copyright and shall appear in the following or similar format: “Reprinted/Reproduced, by permission from diversitydatakids.org. ©2021. Brandeis University”.) The Original Content may not be used or distributed commercially without Brandeis’ express prior written consent.

2.3. Making data interoperable

We will be using the following data formats and types:

- **Comma-separated Values** (CSV)

  A comma-separated values (CSV) file is a delimited text file that uses a comma to separate values. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format. A CSV file typically stores tabular data (numbers and text) in plain text, in which case each line will have the same number of fields.

  It is a standardized format.

2.4. Increase data re-use

As stated already in Section 2.2, all of our data can become completely open over time.

We do not plan to be archiving data (using so-called *cold storage*) for long term preservation already during the project.

To validate the integrity of the results, the following will be done:

- We will run part of the data set repeatedly to catch unexpected changes in results.

3. Other research outputs

We use Data Stewardship Wizard for planning our data management and creating this DMP. The management and planning of other research outputs is done separately and is included as appendix to this DMP. Still, we benefit from data stewardship guidance (e.g. FAIR principles, openness, or security) and it is reflected in our plans with respect to other research outputs.

4. Allocation of resources

FAIR is a central part of our data management; it is considered at every decision in our data management plan. We use the FAIR data process ourselves to make our use of the data as efficient as possible. Making our data FAIR is therefore not a cost that can be separated from the rest of the project.

None of the used repositories charge for their services.

We have a reserved budget for the time and effort it will take to prepare the data for publication. For making data or other research outputs FAIR, we budgeted: ~5 hours.

Jennifer is responsible for reviewing, enhancing, cleaning, or standardizing metadata and the associated data submitted for storage, use and maintenance within a data centre or repository.

Jennifer is responsible for maintaining the finished resource.

To execute the DMP, no additional specialist expertise is required.

We do not require any hardware or software in addition to what is usually available in the institute.

5. Data security

Project members will not store data or software on computers in the lab or external hard drives connected to those computers. They will not carry data with them (e.g. on laptops, USB sticks, or other external media). All project web services are addressed via secure HTTP (https://...). Project members have been instructed about both generic and specific risks to the project.

The possible impact to the project or organization if information is lost is small. The possible impact to the project or organization if information is leaked is small. The possible impact to the project or organization if information is vandalised is small.

We are not using any personal information.

We are not running the project in a collaboration between different groups nor institutes. Therefore, no collaboration agreement related to data access is needed.

6. Ethics

Data we collect

We will not collect any data connected to a person, i.e. "personal data".

The data collection is not subject to ethical legislation.

7. Other issues

We use the Data Stewardship Wizard with its *Common DSW Knowledge Model* (ID: dsw:root:2.6.2) knowledge model to make our DMP. More specifically, we use the https://researchers.ds-wizard.org/wizard DSW instance where the project has direct URL: https://researchers.ds-wizard.org/wizard/projects/c7f3850c-95a8-40ec-acc8-5a491ec47e7e.

We will not be using any extra national, funder, sectorial, nor departmental policies or procedures for data management.