Mapineq

Data Management Plan (DMP)

Mijail Figueroa González
University of Turku

Mapineq deliverable D9.3 March 2023

Funded by
the European Union

Suggested citation:

Figueroa González, M. (2023). Data Management Plan (DMP). Mapineq deliverables. Turku: INVEST Research Flagship Centre / University of Turku. DOI: 10.5281/zenodo.10400966

Summary history		
Version	Date	Comments
1.0	21.02.2023	First draft for internal check
2.1	03.03.2023	Complete draft submitted to UTU library
2.2	22.03.2023	Version delivered to the European Commission
2.3	15.08.2023	Minor mistakes corrected
2.4	18.12.2023	DOI from Zenodo

Mapineq – Mapping inequalities through the life course– is a three-year project (2022-2025) that studies the trends and drivers of intergenerational, educational, labour market, and health inequalities over the life course during the last decades. The research is run by a consortium of eight partners: University of Turku, University of Groningen, National Distance Education University, WZB Berlin Social Science Center, Stockholm University, Tallinn University, Max Planck Gesellschaft (Population Europe), and University of Oxford

Website: www.mapineq.eu

The Mapineq project has received funding from the European Union's Horizon Europe research and innovation programme under the grant agreement No. 101061645.

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union, the European Research Executive Agency, or their affiliated institutions. Neither the European Union nor the granting authority can be held responsible for them.

Acknowledgement:

This document was reviewed by the University of Turku library services as part of Mapineq quality assurance procedures. The content of the document, including opinions expressed and any remaining errors, is the responsibility of the authors.

Publication information:

Content

Tables

Funded by the European Union

Data Management Plan (DMP)

Mapineq will compile, curate, and distribute an open-access inequality database. This database of local to national geo-located indicators will be linked to several survey and register data to provide a more fine-grained approach to the study of social inequalities. Mapineq will also investigate citizens' perceptions on inequalities and important policy related actions that should be addressed in local decision making. The data management plan (DMP) outlines procedures and strategies for managing data through the lifecycle of Mapineq and beyond, and to maximise the impact of the project.

The Mapineq inequality database will be openly accessible through the visualisation tool on the website, and an API for data analysts to query the database programmatically

The source code will be publicly available and will include the data processing steps required to reformat source data for the database

The DMP is a living document that can be updated several times along the lifetime of the project

1. Data summary

1.1. Mapineq inequality database

The project compiles, curates, and distributes its own Mapineq inequality database. The Annex A offers a non-exhaustive list of the original datasets from where the different geo-located indicators might be extracted. The list will be updated periodically as new sources are made available and/or new sources are encountered.

Purpose: The Mapineq inequality database includes institutional and policy measures (e.g., education, family, labour market, social benefits, health, tax related), physical environment (e.g., pollution, green space) and innovative socio-economic indicators from commercial companies and unconventional sources (e.g., house/rental prices, consumer and digital behaviour, perceptions of inequality). This database will be geo-linked from local to national, across time, birth cohort and with key socio-demographic life course measures. The Mapineq inequality database will allow to link the local, regional and national indicators to micro-level datasets in the other WPs, to assess the role of such indicators on inequality trends.

Expected size: > 100GB.

1.2. Re-used data

The Mapineq project re-uses several ready-made data (e.g., secondary data) collected and distributed by either European or national research infrastructures, and population registries.

Purpose: the different secondary datasets and population registries are of specific relevance for different work packages/tasks (see table below) in the project. The link between these micro-data and the local to national geo-located indicators from the Mapineq inequality database will allow to offer a finer-grained view about the trends of inequalities over the life course, and to discover new drivers of social inequalities.

Table 1. Survey data used in Mapineq

Short name	Long name	Data provider & usage license	WPs	Tasks
GGS	Gender and Generation Survey	Generations & Gender Programme https://www.ggp-i.org/data/ggp-research-ethics/	WP2	T2.1
			WP3	T3.1
SOEP	German Socio-Economic Panel	DIW Berlin, the German Institute for Economic Research https://www.diw.de/en/diw_01.c.601584.en/data_access.html	WP2	T2.2, T2.3, T2.5
			WP5	T5.1, T5.2, T5.3, T5.4
BHPS	British Household Panel Survey	Institute for Social and Economic Research, University of Essex https://www.iser.essex.ac.uk/bhps/acquiring-the-data https://ukdataservice.ac.uk/app/uploads/cd137-enduserlicence.pdf	WP2	T2.2, T2.3, T2.5
			WP5	T5.1, T5.2, T5.3, T5.4
UKHLS	UK household Longitudinal Study	Institute for Social and Economic Research, University of Essex https://www.understandingsociety.ac.uk/documentation/access-data https://ukdataservice.ac.uk/app/uploads/cd137-enduserlicence.pdf	WP2	T2.2, T2.3, T2.5
			WP5	T5.1, T5.2, T5.3, T5.4
SHP	Swiss Household Panel	FORS - Swiss Centre of Expertise in the Social Sciences https://www.swissubase.ch/en/catalogue/studies/6097/17007/datasets/932/2435/contract/usage-license	WP2	T2.2, T2.3, T2.5
			WP5	T5.1, T5.2, T5.3, T5.4
ESS	European Social Survey	European Social Survey European Research Infrastructure (ESS ERIC) https://www.europeansocialsurvey.org/data/conditions_of_use.html	WP3	T3.1, T3.2.
			WP7	T7.2
PISA	Programme for International Student Assessment	Organisation for Economic Co-operation and Development https://www.oecd.org/termsandconditions/	WP3	T3.4

FPE	French Panel d'Élèves du Second Degré	Directorate of Evaluation, Forecasting and Performance Monitoring (DEPP), Ministry of National Education and Youth https://data.progedo.fr/studies/doi/10.13144/lil-0955?tab=access	WP3	T3.5
NEPS	National Education Panel Study	Leibniz Institute for Educational Trajectories (LIfBi) https://www.neps-data.de/Data-Center/Data-Access https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Datenzugangswege/Vertraege/NEPS_DataUseAgreement_en.pdf	WP3	T3.5
			WP4	T4.4
EU-SILC	European Union Statistics on Income and Living Conditions	Eurostat https://ec.europa.eu/eurostat/documents/203647/771732/How_to_apply_for_microdata_access.pdf/82d98876-75e5-49f3-950a-d56cec15b896 https://ec.europa.eu/eurostat/documents/203647/771732/Individual_Confidentiality_Declaration.pdf https://ec.europa.eu/eurostat/documents/203647/771732/Confidentiality_undertaking_Terms_of_use.pdf	WP4	T4.1
			WP5	T5.1, T5.2, T5.3
			WP6	T6.2
			WP5	T5.1, T5.2, T5.3
EU-LFS	European Union Labor Force Survey	Eurostat https://ec.europa.eu/eurostat/documents/203647/771732/How_to_apply_for_microdata_access.pdf https://ec.europa.eu/eurostat/documents/203647/771732/Individual_Confidentiality_Declaration.pdf https://ec.europa.eu/eurostat/documents/203647/771732/Confidentiality_undertaking_Terms_of_use.pdf	WP4	T4.2 T4.3
			WP5	T5.1, T5.2, T5.3
			WP6	T6.3
SHARE	Survey of Health, Ageing and Retirement in Europe	SHARE BERLIN Institute https://share-eric.eu/data/data-access/conditions-of-use https://share-eric.eu/fileadmin/user_upload/SHARE_Data_Statement.pdf	WP6	T6.4
INVALSI	Istituto nazionale per la valutazione del sistema educativo di istruzione e formazione	INVALSI Statistical Service https://invalsi-serviziostatistico.cineca.it/	WP3	T3.5
AH	The National Longitudinal Study of Adolescent to Adult Health	Add Health https://addhealth.cpc.unc.edu/data/	WP3	T3.5
HRS	Health and Retirement Study	University of Michigan https://hrs.isr.umich.edu/data-products?_ga=2.13562679.1632109878.1677574858-994909340.1677574858	WP3	T3.5

WLS	Wisconsin Longitudinal Study	University of Wisconsin–Madison https://researchers.wls.wisc.edu/data/survey-data/	WP3	T3.5

Table 2. Population register data used in Mapineq

Country	Data provider & usage license	WPs	Tasks
Finland	Statistics Finland https://www.stat.fi/meta/tietosuoja/kayttolupa_en.html	WP2	T2.2, T2.3, T2.3, T2.4, T2.5
		WP5	T5.1, T5.2, T5.3, T5.4
Sweden	Statistics Sweden https://scb.se/en/services/ordering-data-and-statistics/ordering-microdata/mona--statistics-swedens-platform-for-access-to-microdata/rules-and-regulations/terms-of-use/ https://scb.se/en/services/ordering-data-and-statistics/ordering-microdata/	WP5	T5.1, T5.2, T5.3, T5.4
United Kingdom	Office for National Statistics https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/secureresearchservice https://www.adruk.org/data-access/data-access/	WP2	T2.4
Estonia	Statistics Estonia https://www.stat.ee/sites/default/files/2021-07/Procedure%20for%20the%20dissemination%20of%20confidential%20data%20for%20scientific%20purposes_EN.pdf	WP4	T4.4

1.3. Qualitative data

The project will also collect qualitative data from four online citizen panels conducted in Finland, Germany, Spain and the United Kingdom, from qualitative interviews to stakeholders, and through participant observation. Qualitative data collected through the citizen panels consist of written input from registered participants on the INVEST Next Generation (InNEXT) Co-creation Platform. Qualitative data collected through interviews consist in voice recordings. Qualitative data from participant observation consist in text notes.

Purpose: the citizen panels seek to capture citizens' perceptions on inequalities. It will enable us to engage citizens and explore they challenges they experience and find out what citizens perceive as important policy related actions that should be addressed in local decision making. Additionally, through these panels we want to find out what inequality indicators citizens think should be included in the MapIneq database mapping inequality landscapes. The materials produced will be used as input in the Reality Check Workshops and in designing the MapIneq database and MapIneq visualization tool.

All qualitative data will be used in WP7 under task 7.3.

Expected size of the data: Text files < 500MB. MP· files for voice recordings 600MB/hour.

1.4. Data utility

Re-used data. Other social scientist will profit from the open access replication codes of all research published along the project, in this way we assure that the scientific community can make use of the same secondary data either to ease reusability or to improve reproducibility.

Mapineq inequality database. The main utility of the data is for the work packages across the project. The wider science community will benefit of the open access coding tools that will allow to match the database to other micro data sets used in the project. Policy makers and citizens will benefit of the data as the project will provide an open access easy-to-use visualisation tool through a dashboard hosted in the project's website. Furthermore, the visualisation tool will be developed using a feedback loop with researchers and policy makers to allow them to select and download data and engage in visualisation of geographic and temporal trends using interactive maps.

Qualitative data. The main utility of the data is for researchers in the project, particularly in WP1 and WP7. Nevertheless, the data collected from the citizen panels will also be useful for stakeholders and policy makers to get in contact with citizens' perceptions of inequalities in the Reality Check Workshops (WP8).

2. FAIR data

2.1. Making data findable, including provisions for metadata

2.1.1. Persistent identifiers

The Mapineq inequality database (as much as the licensing of the original data allows it) and all of its metadata will be identified by a persistent identifier (DOIs) to promote findability. The source code will be published in GitHub code repository with a DOI from Zenodo. This will include the catalogue of data sources with full citations. The source code will include the data processing steps required to reformat source data for the database. The goal is to provide code and metadata needed to rebuild our database without needing to permanently publish a large volume of data.

The harmonisation of regional codes that we produce for the datasets included in the Comparative Panel File (CPF) as well as the code that we produce to harmonise register data with the CPF will also be provided to the wider research community via GitHub and the CPF website.

Re-used data already comes with persistent identifiers (DOIs), all publications and metadata produced in the project will make reference to the DOIs of the re-used data.

Qualitative raw data (i.e., written input from participants on the InNEXT co-creation Platform, and voice recordings) will only be available to researchers affiliated to the project, and thus will not have a persistent identifier.

2.1.2. Metadata

For data that we produce, we will include descriptive metadata (i.e. documentation and source code), administrative (i.e. data license), and structural (documentation on internal organisation of the data, and even full tutorials for some of our data and API). For external

data that we are redistributing, we may include all three types of metadata, or we may refer back to the original source for structural metadata where appropriate.

We will use GitHub for the source code and metadata with a DOI to make it citeable and findable. We will also be creating a web server which will be discoverable from general search engines, like Google.

For re-used data, Table 3 and Table 4 display the location of metadata facilitated by data providers.

Table 3. Metadata location of survey data used in Mapineq

Data	Metadata location
GGS	https://ggp.colectica.org/
SOEP	https://www.diw.de/en/diw_01.c.789785.en/documentation_of_soep-core__soep_survey_papers.html
BHPS	https://www.iser.essex.ac.uk/bhps/documentation
UKHLS	https://www.understandingsociety.ac.uk/documentation/data-releases
SHP	https://www.swissubase.ch/en/catalogue/studies/6097/17007/files
ESS	https://www.europeansocialsurvey.org/data/
PISA	https://www.oecd.org/pisa/data/
FPE	https://data.progedo.fr/studies/doi/10.13144/lil-0955?tab=documents
NEPS	https://www.neps-data.de/Data-Center/Data-and-Documentation
EU-SILC	https://www.gesis.org/en/missy/materials/EU-SILC/setups
EU-LFS	https://www.gesis.org/en/missy/metadata/EU-LFS/
SHARE	https://share-eric.eu/data/data-documentation
INVALSI	https://www.invalsi.it/amm_trasp/ss_altricontenuti.php?sezione=Accessibilit%E0%20e%20Catalogo%20dei%20dati,%20metadati%20e%20banche%20dati
AH	https://addhealth.cpc.unc.edu/documentation/
HRS	https://hrs.isr.umich.edu/documentation
WLS	https://researchers.wls.wisc.edu/documentation/

Table 4. Metadata location of register data used in Mapineq

Country	Metadata location
Finland	https://taika.stat.fi/en/index
Sweden	https://www.scb.se/en/documentation/
Estonia	https://www.stat.ee/en/submit-data/questionnaires
United Kingdom	https://www.adruk.org/data-access/

2.1.3. General standards followed

The project follows standards of the Data Documentation Initiative (DDI) when producing metadata.

2.2. Making data accessible

2.2.1. Mapineq inequality database

Openness is a central principle in our project's research activities and relates to good scientific practice and is one of the basic values we rely on in our research.

The Mapineq inequality database will be maintained, curated, and distributed open access through the Leverhulme Centre for Demographic Science (LCDS) for at least 5 years after the end of the project. The database infrastructure will be built using free and open access software such as PostgreSQL, PostGIS, Docker, Python and Flask. The code developed for the MapIneq database will be made openly available through a public GitHub repository and registered with a DOI to make it permanently accessible.

The Mapineq inequality database is accessible in two ways.

1. Website: This will allow users to navigate an interactive map to select data geographically and to filter results by characteristics of data in those locations. This will allow users to easily download subsets of the data that are relevant to them without needing to download the entire database.
2. API: This will allow data analysts and web developers to query the database programmatically using the R Statistical Programming language, Python, JavaScript (e.g. our own website will use this mechanism), or any other programming language that supports API requests (i.e. http GET or POST requests).

We anticipate that the database will be hundreds of gigabytes in size or larger. Therefore, it is unlikely to be feasible for users to download the entire database. This is why it is important to allow users mechanisms to conveniently filter data to download subsets (i.e. the website and API).

We do not plan to include any personal or otherwise sensitive data in the public facing MapIneq database, and so we do not anticipate a need to enforce user authentication.

2.2.2. Re-used data

Re-used data is accessible at data providers websites. However, all the coding tools and metadata produced around the project will be made open access available under the Creative Commons Attribution-Non-Commercial-Share Alike 4.0 International (CC BY-NC-SA 4.0) License or similar and published in recognised repositories such as GitHub.

2.2.3. Qualitative data

The qualitative raw data (i.e., text input form participants in citizen panels, voice recordings, and text notes from participant observation) will only be accessible to researchers. Qualitative raw data from citizen panels will be pseudonymised by erasing any references to personal data such as, names, and other personal information. Pseudonymised text without references to personal data will prevent for re-identification of participants. Pseudonymised text will be translated to English. Pseudonymised text will then be used as an input for the Reality Workshops with stakeholders and policy makers.

Other qualitative data (i.e., voice recordings, and text notes from participant observation) will not be made publicly available.

Qualitative data will be stored for five years after the projects ends, after five years of storage the data will be permanently destroyed.

2.3. Making data interoperable

Interoperability of the data and metadata produced in the project will be assured by following standards best practices commonly used in the social sciences.

We will use data formats that are interoperable across operating systems and that can be accessed using free open-source software.

Data and metadata will be assigned with terms coming from established vocabularies such as:

1. The European Language Social Science Thesaurus
2. The Sociology Vocabulary

Furthermore, we commit to follow the best practices for interoperability in the social sciences according to the Data Documentation Initiative (DDI, https://ddialliance.org), and the Consortium of European Social Science Data Archives (CESSDA, https://www.cessda.eu).

2.4. Increase data re-use

The Mapineq inequality database will be openly accessible for anyone. We will follow FAIR practices in the development and publication of the Mapineq database source code. We will make our code and data findable using a permanent citeable DOI. We will make it accessible through an open access GitHub code repository and publicly accessible web servers. This will include complete citations to source data and their DOIs (where available). Our coding practices will emphasize interoperability by using free open-source software that runs on Windows, Apple, and Unix-based operating systems. Our open source and permanently published code will ensure re-usability even after the time horizon for the project and maintenance of web servers for the Mapineq website, API, and SQL database. The open code base will allow future users to rebuild and modify these resources using our source code.

The re-used survey data are owned by data providers. The register data are owned by the statistical offices who compile them from the register sources. We will provide open access to replication files for all research published in the project via GitHub. The replication files will include references to the persistent identifiers of the re-used data, and code for data cleaning, data linkage, and analysis.

Data and metadata will be published under the Creative Commons Attribution-Non-Commercial-Share Alike 4.0 International (CC BY-NC-SA 4.0) license or similar.

2.4.1. Quality assurance

The Mapineq inequality database quality will be maintained by the Leverhulme Centre for Demographic Science (LCDS) for at least 5 years after the end of the project.

For the re-used survey data, the data providers do consistent checks for the quality and release corrected versions of the data, if needed. Researchers of the project will used the most recent versions published by data providers.

Register data are compiled from administrative register records by the responsible statistical agencies. Any corrections to the raw data are done at all times by the statistical offices themselves, both if pointed out by the administrative users of the original data or by the scientific users of compiled datasets.

For the qualitative data gathered from the citizen panels, quality is assured by the involvement of highly qualified researchers as moderators in the citizen panels. Moderators will have excellent language skills of the language of the country in which the panel takes place. Quality is also assured by the use of the INVEST Next Generation (InNEXT) Co-creation Platform for collecting the data.

Qualitative interviews will be conducted by experienced researchers. Semi-structured interview guides will be developed and evaluated before the actual interviews.

3. Other research outputs

The project does not expect any other research outputs than those described in the previous sections.

4. Allocation of resources

4.1. Mapineq inequality database

The cost of compiling, curating, and making openly accessible the database and its metadata will be covered by the project. Under WP1 Inequality database, the beneficiary partner, University of Groningen, will employ three researchers for this endeavour: two full-time and one part-time. Additionally, the associated partner, University of Oxford, will employ two researchers with their own funds.

The project will make use of GitHub and Zenodo which are free services that provide long-term repositories of code and metadata. There will be, however, costs for the web server (Amazon Web Services) which will provide the mechanism for our live website, API, and database. We have estimated this to cost about $1000 USD per year. This could go higher if we need exceptionally large storage capacity.

The database will be maintained at least 5 years after the end of the project by the Leverhulme Centre for Demographic Science (LCDS). The database and the metadata will be available in the month 24 of the project, as stated in the Gran agreement, Annex 1.

The data controller for the Mapineq inequality database is Dr. Douglas Leasure (University of Oxford/Leverhulme Centre for Demographic Science).

4.2. Re-used data

Re-used data are available through the internet.

Researchers employed fully or partially through the project's funds will make their coding tools openly available as part of their usual research practices. The INVEST Research Flagship Centre at the University of Turku (partner coordinator) has estimated that researchers working with re-used data commonly have a time usage ratio for coding vs.

other parts of research of 5:1. In theory the code can be made to a shareable version within the same time frame. This ratio is 10:1 for researchers working with register data.

Each WP leader is responsible for the management of the re-used data used within the work package they lead. The management of the data by WP leaders and contributors should respect the terms of the original providers (see section 1.2), and in particular what is specified in section 5 about data security.

4.3. Qualitative data

Qualitative data will not be openly accessible. Nevertheless, Researcher Dr. Hanna Ylöstalo, from the University of Turku, is the data controller for the qualitative data.

5. Data security

5.1. Mapineq inequality database

The MapIneq database will only include aggregated anonymized data that have data licenses allowing redistribution. It will be hosted on a Linux-based server within the European Union (e.g. Amazon Web Services). The Linux servers will use standard security protocols such as regular software updates/upgrades and native firewall software (e.g. ufw) to prevent unauthorized access. Public access to Mapineq database will only be possible through the API and web applications that use the API. This will provide full access to the data while preventing direct access to the SQL database which reduces security risks to the server. Standard firewall rules and PostgreSQL user controls will be implemented to restrict direct access to the Mapineq database to the project team and collaborators.

5.2. Re-used data

Once received by data providers, ready-used data, survey and population registries, are kept stored in secure physical or virtual storage with depending on the requirement form data providers. Each partner will make use of the secured storage infrastructure offered by their own institutions following the requirements of the different data.

5.2.1. GGS

Data can only be used exclusively by researchers affiliated with the Organization who have signed the Pledge of Confidentiality and returned it to the Co-ordination Centre. Data will be stored in a safe location. Any hard copies will be kept in locked storage, while the data in digital format shall be kept on computers protected with passwords with no open access to the internet.

5.2.2. SOEP

Only the individuals named in the contract may work with the data, data can be stored locally, however, only on a computer where only the researchers with valid user permissions can have access to. Every member who will be using the SOEP data must agree to adhere to the data protection regulations. All users are responsible for destroying or deleting the SOEP data (including all backup copies, auxiliary files, and dump files) when

they terminate their data distribution contract and/or leave the institution where the contract was signed so that no one else can work with the data.

5.2.3. BHPS & UKHLS

The means of access to the data (such as passwords) are kept secure and not disclosed to a third party except by special written permission or licence obtained from the original Data Service Provider. At the end of the access period, we're abide to destroy all copies of the data, including temporary copies, printed copies, personal copies, back-ups, subsets of variables/cases, derived datasets and all electronic copies including copies held on portable media.

5.2.4. SHP

Data cannot be transmitted to third parties, whether in original or modified form. Data must be store in a way such that no third party can gain access to them. Data must be destroyed at the latest upon expiry of the present contract, data users have to confirm this to FORS.

5.2.5. ESS

The data are available without restrictions, for not-for-profit purposes. Data is distributed under CC BY-NC-SA 4.0. Users are free to share — copy and redistribute the material in any medium or format and adapt — remix, transform, and build upon the material.

5.2.6. PISA

The data can be copied, adapted, printed, distributed, shared and embedded for any purpose, even for commercial use. Researchers must give appropriate credit to the OECD.

5.2.7. FPE

Users are responsible for protecting the confidentiality of the data, and to destroying the files once the research work is completed. Users are also not allowed to transfer the data to any third party.

5.2.8. NEPS

Only researchers listed in the NEPS Data Use Agreement can access the data. The data recipient and the persons involved in the research project will ensure that the data, including those in modified form, possible backup copies, excerpt and/or auxiliary files, will be deleted on all data carriers where they are stored when the agreement expires and/or terminates. The Federal State label in the starting cohorts of schools and higher education institutions can only be accessed via remote access (RemoteNEPS) and by logging in as a guest in Bamberg (On-site).

5.2.9. EU-SILC & EU-LFS

Data must be stored on a password-protected computer. Access to the confidential data must be restricted to authorised researchers named in the research proposal. After completion of the project, the principal researcher must destroy the confidential dataset provided by Eurostat and any confidential data derived from it and sign a declaration to the effect that it has been ensured that all confidential data have been destroyed.

5.2.10. SHARE

The copies of the SHARE data downloaded from the SHARE Research Data Center will be protected on a password-protected computer only available to researchers that have signed the conditions of use. Users are not allowed to make copies of the data available to others and/or enable any third-party access to the database.

5.2.11. INVALSI

data with the geo-localization of schools have been achieved through a special agreement with INVALSI and cannot be transmitted to third parties, whether in original or modified form.

5.2.12. Add Health

The Add Health data is available in two forms, public-use data and restricted-use data. To have the restricted-use data files require a restricted-use contract. To be eligible to enter into a contract, researchers must complete a contract application which includes: security plan, IRB approval letter, $1000 payment by check, and sign a data-use contract agreeing to keep the data confidential. There are three options to store the data. Data can be stored on an 1) Encrypted Stand-Alone Desktop Computer, 2) Encrypted External Hard Drive, 3) Remote Compute Server. The security protocols are different accord to the options decided. To have access to the genetic data is necessary to apply through the NIH database of Genotypes and Phenotypes (dbGaP) and then apply and pay Add Health to have the link Add Health data file. Therefore, regarding the genetic data it is necessary to follow the dbGaP conditions of use.

5.2.13. HRS

Once the researchers have registered as a user of the HRS site, public release files may be downloaded to own computer from their private server. By registering for access to HRS Public Release data, the user agrees to all of the following. a) Make no attempts to identify study participants; b) Not to transfer HRS Public Release data to any third party other than staff or students for whom you are directly responsible except as indicated below; c) Not to allow others to use your username and password to access this site; d) To certify the destruction of any downloaded Public Release data file as well as any data files derived from the downloaded file when requested to do so by the Health and Retirement Study. Since the HRS genetic data are stored in the the database of Genotypes and Phenotypes (dbGaP) system, to have the genetic data is necessary to apply through dbGaP and therefore it is necessary to follow the dbGaP conditions of use.

5.2.14. WLS

Public data access requires a registration on the WLS web site before downloading the data. There are protected measures, which are those measures that increase a subject's risk of identifiability, and they are available to researchers by request. Researchers wanting to use the genetic data may submit a proposal that will be reviewed by an independent board. The board reviews the proposals to ensure that the confidentiality of participants' information will be protected. Moreover, researchers have to submit

statement acknowledging the potential pitfalls of genomics research and positioning themself against overly reductive or deterministic uses of the genomic data.

5.2.15. Finnish register data

The Finnish register data 1 (running 1970-2007) is stored on secured computers in the locked rooms of University of Turku. The access to the data is allowed only to the persons who have applied and have been provided user permission by Statistics Finland. The Finnish register dataset 2 (running 1980-2010) are not allowed to travel but will stay in the highly secure database of Statistics Finland, only allowing the researchers with the user permission to remotely access these using high security internet connections. The remote systems can be only accessed through the university networks in question. The researchers are not allowed to access raw data or personal identification numbers and combinations with other datasets are not allowed.

5.2.16. Swedish register data

Swedish register data are mainly used on Statistics Sweden's MONA (Microdata Online Access) platform where licensed users process data through a secure connection without the data leaving Statistics Sweden. Microdata to be used for research is applied for through Statistics Sweden. Registers held by other government agencies are in general delivered to Statistics Sweden where, subject to data confidentiality and approval by the relevant authorities, the registers are combined for research use.

5.2.17. Estonian register data

The Estonian register data is not allowed to travel but will stay in the highly secure database of Statistics Estonia, only allowing the researchers with the user permission to remotely access these using high security internet connections. The remote systems can be only accessed through the ID verification and VPN. The researchers are not allowed to access raw data or personal identification numbers and combinations with other datasets are not possible to do by researchers.

5.2.18. United Kingdom register data

Register data is accessible to researchers based in organisations that have an Assured Organisational Connectivity (AOC) agreement in place through the Secure Research Service (SRS) from their employers' offices. Access must be from a machine provided by the organisation the researcher works for, connecting to the SRS through their corporate virtual private network (VPN), access must be from within the UK, the access to the SRS from outside the UK is not permitted. If researchers access the SRS from home will need to read and sign an extended Security Operations Procedures (SyOps) document.

5.3. Qualitative data

Qualitative raw data (i.e., text input and voice recordings) are pseudonymized after collection. Pseudonymized data and personal data from participants are stored in separate files on secured computers in the locked rooms of the University of Turku. Researcher Hanna Ylöstalo, from the University of Turku, is the data controller for the qualitative data, only she will have access to the files that allow to link the pseudonymized data with the

personal data. This is necessary in case participants require their data or want their data to be removed from the study.

Raw data from citizen panels is collected in four different languages. Translators will only have access to pseudonymised data where any references to personal data have been erased (art. 32 GDPR). Translators will sign a contract that prevents them from keeping any local copies of the data after the translations have been done.

Voice recordings are transcribed, any reference to personal data is removed. Researchers work with pseudonymised transcriptions.

6. Ethics

The project is committed to All European Academies (ALLEA) European Code of Conduct for Research Integrity in all research contexts, and the requirements set by the General Data Protection Regulation (GDPR).

6.1. Mapineq inequality database

The main content of the database relates to geo-located indicators at the local, regional and national level. Any project investigating social inequalities and highlighting geographic areas or sub-populations with disparate outcomes risks stigmatizing or otherwise inadvertently causing some harm to outlier populations. The Mapineq project takes this risk seriously and has a dedicated independent ethics advisor to regularly review these risks and our mitigation strategies. We are committed to presenting results with a positive outlook–such as highlighting “areas of opportunity” to improve outcomes rather than areas with “bad outcomes”. We are committed to engaging with policy makers and the public to use these results to effect positive change, and we will regularly solicit their feedback to ensure that we are not presenting data in a way that undermines that primary objective.

The MapIneq database will include only aggregated and anonymized data with data licenses that allow open redistribution. This minimizes or eliminates risks of personal identification through merging with external data sources or disclosures of new data that are not already publicly available.

Additionally, the consortium will develop a protocol for unexpected findings, this protocol will be made publicly available upon approval from the independent ethics advisor.

6.2. Re-used data

All data providers conduct their own ethical review before providing access to the data. Data providers are located in the European Union and thus they are subject to the GDPR. Data providers have asked consent from each individual respondent (art. 7 GDPR). Re-used data comes anonymised directly from data providers; thus it does not involve personal information about identified individuals. User licenses signed by researchers of the project strictly prevent researchers for trying to re-identify individuals.

According to national legislations on administrative register data from Estonia, Finland and Sweden, individual-level consent is not required for the research use of this information. Each data source, including those that include genetic information require additional

evaluation for usage by the data provider and ethics approval by the University where the research is conducted.

6.3. Qualitative data

All participants will be fully informed about the research project, its purpose, and the use of their data. Before data collection starts, participants provide active voluntary consent. For the citizen panels, the active voluntary consent is retrieved via the registration procedure in the InNEXT co-creation platform. Before any participant can submit written comments, they must read and agree with the privacy statement of the InNEXT co-creation platform.

For the interviews, active voluntary consent is collected in written form, only after the written form is read and signed interviews can take place.

Participants are informed of their right to opt out of the study at any time. Participants have full ownership of their data, they are informed of their right to check their recorded data later if required by themselves, by contacting the data controller Dr. Hanna Ylöstalo, from the University of Turku.

Participants will be provided with information on the role of researchers with whom they are involved through participant observation in fieldwork (e.g. in meetings).

Reference to any citations of data provided by participants must be pseudonymised. No connection of individuals, either directly or implicitly through quotations, will be possible. Participants, informed of their right to remain anonymous, are able to halt the observation at any time.

7. Other issues

The Invest Register Data Hub (InReg) is the infrastructure used to produce novel register-based research databases, mainly applying individual level administrative information. Register data can be combined with information gathered directly from research subjects, only with permission of the data owner. The data may in some cases also cover information based on biological samples. While the legislation on register data allows the reuse of data for research purposes, the data gathered in the datasets are typically highly sensitive. That is why all applications are subject to a review by the ethical board of UTU, THL, Finndata and/or Statistics Finland. In addition, some register maintainers may require their own review process (e.g., data on military conscripts). While different registers are linked together typically with person IDs, these data are always anonymized before data are provided to researchers. Typically, also the remote access system providers (e.g., Statistics Finland) conduct their own review on any data extracted from the systems to make sure that individuals cannot be identified.

References

ALLEA – All European Academies (2017). The European Code of Conduct for Research Integrity. ISBN 978-3-00-055767-5

Creative Commons—Attribution-NonCommercial-ShareAlike 4.0 International—CC BY NC SA 4.0. (n.d.). Retrieved 3 March 2023, from https://creativecommons.org/licenses/by-nc-sa/4.0/

FAIRsharing.org: ELSST; European Language Social Science Thesaurus, DOI: 10.25504/FAIRsharing.acd824, Last Edited: Tuesday, November 22nd 2022, 22:01, Last Editor: allysonlister, Last Accessed: Friday, March 3rd 2023, 13:44

FAIRsharing.org: Sociology Vocabulary, DOI: 10.25504/FAIRsharing.8DrzMv, Last Edited: Wednesday, June 15th 2022, 16:10, Last Editor: delphinedauga, Last Accessed: Friday, March 3rd 2023, 13:25

Annex A. Mapineq inequality database sources

Table 5. Original sources of compiled geo-located indicators

Category	Name	Description	Source url	Spatial extent	Spatial resolution	Time period	Frequency	Terms of use	Access restrictions
surveys	IPUMS-International	Census microdata	https://international.ipums.org/international/	Global	subnational admin units	1960s - present (see here)	10 years	no redistribution	approved users only
surveys	ESS	European Social Survey	https://www.europeansocialsurvey.org/data/	Europe	subnational admin units	2002 - 2018	2 years	open-access CC-BY-NC-SA 4.0	none
surveys	INVALSI	Population data survey on school competencies; students aged 7 and 10	https://www.invalsi.it/invalsi/index.php	Italy	commuting zones (CZ)	2019			
other	GHS-SMOD	Map of built-up areas and settlement types (e.g. urban/rural)	https://ghsl.jrc.ec.europa.eu/ghs_smod2019.php	Global	1km	1975 - 2015	5 years (varies)	open-access	none
other	WorldPop Covariates	Assorted geospatial covariates: Nighttime lights, topo, roads, land cover	ftp://ftp.worldpop.org/GIS/Covariates/Global_2000_2020/	Global	100 m	2000 - 2020 (varies)	annual (varies)	open-access CC-BY-4.0	none
housing	Zoopla	Zoopla property listings and sold price data	https://developer.zoopla.co.uk/home	United Kingdom	buildings	2005 - present	daily	can redistribute in aggregate form(?)	approved users only
housing	OSM Buildings	Building footprints and type (e.g. residential) from Open Street Maps	https://download.geofabrik.de/europe.html	Global, patchy	buildings	current	daily	open-access	none

demographics	WorldPop	Population estimates by age and sex	ftp://ftp.worldpop.org/GIS/AgeSex_structures/Global_2000_2020/	Global	100 m	2000 - 2020	annual	open-access CC-BY-4.0	none
demographics	EuroStat Life Expectancy	Life expectancy by age, sex, and NUTS2 regions	https://ec.europa.eu/eurostat/databrowser/view/demo_r_mlifexp/default/table?lang=en	Europe	subnational admin units (NUTS2)	2015 - 2019	annual	open-access	open
boundaries	NUTS	Sub-national region boundaries used by ESS	https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/nuts	Europe	subnational admin units	2003 - 2021	3 years	open-access, non-commercial, attribution	none
boundaries	GADM	Sub-national administrative boundaries	https://gadm.org/data.html	Global	subnational admin units	current	periodic	permission required for redistribution	open for non-commercial
boundaries	IPUMS GIS Boundaries	Harmonized sub-national administrative boundaries	https://international.ipums.org/international/gis.shtml	Global	subnational admin units	static (not sure of year)	NA	open-access	open
surveys	EU-SILC	European Union Statistics on Income and Living Conditions	https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions	EU27/28 + selected neighboring countries	NUTS-2	2004 - 2020 (not all waves for all countries)	annual	no redistribution	approved users/projects at recognised institutions
surveys	EU-LFS	European Uninion Labour Force Survey	https://ec.europa.eu/eurostat/web/microdata/european-union-labour-force-survey	EU27/28 + selected neighboring countries	NUTS-2	1983 - 2020	quarterly	no redistribution	approved users/projects at recognised institutions
surveys	SHARE	Survey of Health, Ageing and Retirement in Europe	http://www.share-eric.eu/home0.html	Europe (28 w/o UK), Israel	Not known, enquiry sent	2004 - 2020	ca. 2 years	no redistribution	approved users/projects at recognised institutions

surveys	NEPS, SC 4	National Educational Panel Study, Starting Cohort 4	https://www.neps-data.de/Mainpage	Germany	Administrative districts (Kreise), on-site only	2010 - 2019	annual until 2016, then biennial	no redistribution	approved projects at recognised institutions
surveys	PISA	Programme for International Student Assessment	https://www.oecd.org/pisa/data/	OECD+others	depends on the country	2000 – 2022	every 3 years, but 2021 moved to 2022	surveys	PISA
other	DESI	Digital Economy and Society Index	https://digital-strategy.ec.europa.eu/en/policies/desi	EU					
surveys	GGS	Generations and gender surveys	https://www.ggp-i.org/data/browse-the-data/	Europe	subnational admin units (NUTS2)				