End report ‘Historical Sample of the Netherlands (HSN, task 2.3)’ within the ODISSEI Roadmap project

Leo van Toor (CBS), Jennifer Claij-Swart (CBS), Ruben van Gaalen (CBS), Rick Mourits (IISH) & Richard L. Zijdeman (IISH), 2022-11-01, 1.0

1 Overview

The ODISSEI acronym stands for Open Data Infrastructure for Social Science and Economic Innovations. ODISSEI is made up of more than forty member organisations that financially contribute to its development. These member organisations include Social Science Faculties, Economics Faculties, Research Institutes, Public Research Agencies, Statistics Netherlands, and E-Infrastructure providers. The organisations in ODISSEI represent more than 5,000 social science researchers and ODISSEI aims to support their research by providing world class data infrastructure services that increase their access to data, computing tools, expertise, and financial resources.

Through ODISSEI, researchers have access to large-scale, longitudinal data collections as well as innovative and diverse new forms of data. These can be linked to administrative data at Statistics Netherlands (CBS). Combining data from a wide range of sources enables researchers to answer new, exciting, interdisciplinary research questions and to investigate existing questions in novel, new ways.

This is the end report of the task ‘Historical Sample of the Netherlands’ (2.3), which was executed as part of the ODISSEI Roadmap project (NWO grant number 184.035.014) between 2020-07-01 and 2022-06-30. The task leader was Richard Zijdeman. For more information, please contact info@odissei-data.nl.

2 Task highlights

2.1 Initial problem statement and goal

One of the aims of the ODISSEI Observatory is to broaden the types of data that are integrated into ODISSEI. The Observatory will seek to integrate data from three new sources and open up new lines of research. In task 2.3, data from the Historical Sample of the Netherlands (HSN) will be linked with Statistics Netherlands Microdata (Social Statistical Database SSD). In practical terms, this involves the creation of a key for the persistent identifiers of individuals in the HSN and the persistent identifier for the same individual in the Social Statistical Database which is available in ODISSEI via CBS microdata services. Using this link it is possible to trace the outcomes of the individuals in the HSN and their descendants forward to the current day. This will enable links between the historical research conducted using the HSN and contemporary society and the outcomes studied by social scientists. This line of research will also enable stronger links with CLARIAH and the broader field of socio-economic history more generally.

2.2 Intended deliverables at the end of the project

- Proof of concept of linking the HSN Life Course database with SSD
- Enlarging the sample for the critical birth period 1903-1922 (from 0.25 to 0.5% of all births).

2.3 Layman summary

The Netherlands has two key databases to study of social inequalities across the life course and over generations: the Historical Sample of the Netherlands (HSN) for persons born in the Netherlands from 1812 till 1922, and the System of Social statistical Datasets (SSD) with register data on all current inhabitants. We aimed to establish a Proof of Concept linkage of the HSN and SSD, allowing for the historical life trajectories to be linked forward to contemporary outcomes. For the Proof of Concept we tested whether linkage could be established on the basis of the combination of three birth dates (ego, father, mother), date of marriage, and gender. We developed an initial linking strategy which we validated and refined based on non-unique links and deviating information between HSN and SSD. Additionally, we established a link between the written (HSN) and coded (SSD) places of birth, and used this information in the validation process. The revised linking strategy results in linkage of 77% of the linkable HSN records. The stringent validation criteria of the linking steps and evaluation of the linked result appear to indicate that we provide a successful Proof of Concept for the linkage of the HSN and SSD as conducted by Statistics Netherlands.

3 Task end report

3.1 Conclusions

We succeeded in our main task to create and evaluate more than a dozen linkage strategies to connect the HSN and SSD datasets. Moreover, we defined a handful of strategies that we would call successful, meaning that contemporary data in SSD could be enriched with direct links to life course information of ancestors. We also aimed to retrieve information on additional life courses to increase

the sample size of the HSN. We indeed did retrieve a large number of records, but due to COVID-19 restrictions access to archives was hampered in the crucial (early) phase of the project. Currently, the IISG is still processing the retrieved records and will release them in a future release, outside the scope of the project. The project started late as the project initiator and two software engineers went into retirement and it took some time to get up to speed with the HSN data pipeline. Moreover, we moved relevant parts of the software to docker containers allowing for easier access and maintenance.

This Proof of Concept (PoC) shows that it is possible to match historical sources to the SSD without the use of names. The successful linkage relies on the following three rules:

- The combination of gender, birth date, parental birth dates, and marriage dates gives enough information to get unique links with high precision.
- To also guarantee high retrieval, a stepwise linking procedure is required as that links unmatched cases using fewer linking variables.
- The results of each of these iterations can be measured by the share of non-ambiguous links and overlap with secondary variables, such as the place of birth or unused linking variables.

Combined, these three rules pave the way for longitudinal research on 200 years of data and multigenerational research on over 6 generations. Moreover, it shows that similar linking efforts are also possible for datasets that contain information on individuals without a family history in the contemporary Netherlands, such as the former overseas territories or migrants in general.

Given the positive outcome of the PoC, as a next step we would like to see whether we can actually offer PoC data within the SSD system. For that we would need to do two things. One is to make sure that the addition of HSN data to SSD safeguards the anonymity of individuals in SSD. This could be done by providing selections of the data or aggregating data. Specifically, we think of variables such as background SES, family size, longevity of parents and migration pattern. The other is to investigate the representativeness of the PoC for the contemporary Dutch Population to assist researchers in making valid claims about the research outcomes.

Experiences from and results gained in this task were crucial for submitting the NWO Research Infrastructure grant proposal '[MULTIGENS]', in which we propose the linking of HSN and SSD at large. The proposal is currently under review (with the final outcome being expected in 2023).

3.2 Dissemination activities

We reported on our progression during the ODISSEI conferences in 2021 and 2022 and have written a research report, that is freely available via: https://confluence.socialhistoryservices.org/display/HSNDB/PoC+Linking+HSN+with+SSD. As mentioned, we did not release any new data, but records gathered during this task will be part of a new HSN release that will be available via the IISG's dataverse instance: https://datasets.iisg.amsterdam/dataverse/HSNDB-HSN.