Fair and Tender Data. The FAIRness of Four Databases With Historical Individual Life Course Data Tested

By Lex Heerma van Voss

To cite this article: Heerma van Voss, L. (2021). Fair and Tender Data. The FAIRness of Four Databases With Historical Individual Life Course Data Tested. *Historical Life Course Studies*, *10*, 13-18. https://doi.org/10.51964/hlcs9559

HISTORICAL LIFE COURSE STUDIES

Not Like Everybody Else.
Essays in Honor of Kees Mandemakers

VOLUME 10, SPECIAL ISSUE 3
2021

GUEST EDITORS
Hilde Bras
Jan Kok
Richard L. Zijdeman

EHPS
NETWORK

MISSION STATEMENT

HISTORICAL LIFE COURSE STUDIES

Historical Life Course Studies is the electronic journal of the *European Historical Population Samples Network* (EHPS-Net). The journal is the primary publishing outlet for research involved in the conversion of existing European and non-European large historical demographic databases into a common format, the Intermediate Data Structure, and for studies based on these databases. The journal publishes both methodological and substantive research articles.

Methodological Articles
This section includes methodological articles that describe all forms of data handling involving large historical databases, including extensive descriptions of new or existing databases, syntax, algorithms and extraction programs. Authors are encouraged to share their syntaxes, applications and other forms of software presented in their article, if pertinent, on the openjournals website.

Research articles
This section includes substantive articles reporting the results of comparative longitudinal studies that are demographic and historical in nature, and that are based on micro-data from large historical databases.

Historical Life Course Studies is a no-fee double-blind, peer-reviewed open-access journal supported by the European Science Foundation (ESF, http://www.esf.org), the Scientific Research Network of Historical Demography (FWO Flanders, http://www.historicaldemography.be) and the International Institute of Social History Amsterdam (IISH, http://socialhistory.org/). Manuscripts are reviewed by the editors, members of the editorial and scientific boards, and by external reviewers. All journal content is freely available on the internet at https://openjournals.nl/index.php/hlcs.

Co-Editors-In-Chief:
Paul Puschmann (Radboud University) & Luciana Quaranta (Lund University)
hislives@kuleuven.be

The European Science Foundation (ESF) provides a platform for its Member Organisations to advance science and explore new directions for research at the European level. Established in 1974 as an independent non-governmental organisation, the ESF currently serves 78 Member Organisations across 30 countries. EHPS-Net is an ESF Research Networking Programme.

The European Historical Population Samples Network (EHPS-net) brings together scholars to create a common format for databases containing non-aggregated information on persons, families and households. The aim is to form an integrated and joint interface between many European and non-European databases to stimulate comparative research on the micro-level.
Visit: http://www.ehps-net.eu.

HISTORICAL LIFE COURSE STUDIES
VOLUME 10, SPECIAL ISSUE 3 (2021), 13-18, published 31-03-2021

Fair and Tender Data

The FAIRness of Four Databases With Historical Individual Life Course Data Tested

Lex Heerma van Voss

Huygens Institute for the History of the Netherlands & Utrecht University

ABSTRACT

Four databases with data on individual historical life courses are tested for FAIRness: the TRA, Umeå, HSN and IPUMS databases. All databases make their data much more Findable than they were in the original sources. But as databases, they are best findable if their name is a unique acronym, and if different sub-datasets all use that same acronym. Sensitive data have to be protected. Two databases make anonymous data sets or those only containing information on deceased individuals Accessible without any formalities, and other databases could follow this example. To increase Interoperability a large number of tools are offered by the databases. Reusability is among the *raisons d'être* of these databases.

Keywords: FAIR data, Historical life course databases, Findable, Accessible, Interoperable; Reusable

e-ISSN: 2352-6343
DOI article: https://doi.org/10.51964/hlcs9559
The article can be downloaded from here.

Come all ye fair and tender ladies
Take warning how you court your man
They're like a star on a summer morning
First they appear then they're gone again

1 INTRODUCTION

A traditional Appalachian folk song, recorded by numerous folk and country artists, tells fair and tender ladies to take care in love. Men are like a heavenly body, which is visible before dawn but disappears from sight at sunrise. In other versions of the song, the faithless male lovers are likened to sparrows or swallows flittering away (Roud no. 451). Even if the lover has disappeared from sight, the love may have left visible traces, in life and in historical records. The latter are in turn chased by historians. While historical individuals may seem as fleeting as morning stars or birds, historians come to grips with their life courses by collecting life events in historical databases. These allow us to describe collective life courses over generations, and to contrast singular or group experiences with the average life course in historical societies. Quaranta (2015) gives an overview of the kind of questions such data allow us to answer.

Data on individual historical life courses are collected in dozens of major databases accessible online.[1] There is broad agreement that such digital assets should be FAIR: Findable, Accessible, Interoperable, and Reusable (www.go-fair.org/fair-principles/). The following represents a very simple test of the FAIRness of a few of the larger databases which offer online demographical data on historical life courses. Approaching the databases online as a naïve user, the fairness of the databases is assessed from the position of the outsider, aiming to access the data. Being interoperable and reusable is more or less a *raison d'être* of this type of database, which is not saying that these do not pose their challenges. But the findability and accessibility of the data will get most of our attention.

For this assessment we selected four high profile databases. The TRA survey aims at life reconstitution of all individuals whose last name begins with the letters 'Tra' and their descendants who died in France between 1800 and 1939. It contains data from population registers, military and fiscal records (Bourdieu, Kesztenbaum, & Postel-Vinay, 2014).[2] The Umeå database is based on the historical parish records of northern Sweden, from the 18th and the 19th centuries, and extended for a subset to the 1950s (Edvinsson & Engberg, 2020).[3] The HSN (Historical Sample of the Netherlands) contains life courses based on population registers and additional sources for a 0.5% sample of those born in the Netherlands between 1812 and 1922 (Mandemakers & Kok, 2020).[4] IPUMS contains data on individual and household level from the U.S. decennial censuses from 1790 to 2010 (Ruggles et al., 2020).[5]

The first question is how *findable* the databases are. This simple question was approached in a simple way. Each database was both searched for under the name it seems to be known for in the historical profession and as 'database historical demography <country name>'. Both searches were executed in Google Chrome in the first two weeks of February 2021.

2 FINDABLE

The TRA database may be known as such in the historical profession, but looking for it as 'TRA data set' only gives one relevant hit among the first 50 results. It is at place thirteen and refers to an article on researchgate.net on parental status homogamy in France in the 19th century. 'TRA database' gives a hit at place eight.

[1] Overviews are for instance available at https://ehps-net.eu/databases, https://international.ipums.org/international/index.shtml and https://censusmosaic.demog.berkeley.edu/data/mosaic-data-files (consulted 14 February 2021).
[2] https://tra.site.ined.fr/en/
[3] https://www.umu.se/en/centre-for-demographic-and-ageing-research/databases/
[4] https://iisg.amsterdam/nl/hsn
[5] https://ipums.org/

Searching for the database under its official name 'l'Enquête TRA' leads to a page of relevant hits, topped off by the official website of the project. However, to get this satisfactory result requires awareness of this rather specific official name. Somewhat surprisingly, the search for 'database historical demography France' leads to better results. The second hit is to the European Historical Population Samples Network website, which list the database as 'Base TRA Patrimoine' and links to the official website.

The Demographic Database at Umeå University is easily found under this precise name or variants of it. 'Database historical demography Sweden' has a hit on the first page of results, at place seven. 'Database historical demography Umeå' and 'Swedish Historical Population Statistics' give the database as the first hit, but 'Sweden Historical Population Statistics' only as the twelfth hit, and then only indirectly. For non-Swedish historians working with similar data the name of the university town may be strongly linked to this specific historical demography database, but outside this circle the name of a large Swedish town and the sixth university of the country must have more associations. At Umeå, the Demographic Database is part of the Centre for Demographic and Ageing Research (CEDAR). It houses the data sets POPUM and POPLINK, based on parish registers, TABVERK, which contains data on the population of Swedish parishes from 1749 to 1859, as reported by the clergymen to *Tabellkommissionen* in Stockholm (*Tabellverket)* and FOLKNET, which contains local and regional data derived from Statistics Sweden. There are good reasons to employ all these different names. TABVERK refers to the historical Tabellverket, and FOLKNET was originally collected by an individual scholar, and it makes sense to record these as separate entities. But all in all it is less than clear to the uninitiated visitor what exactly the Demographic Database is. And if one searches, for instance, for Poplink, that acronym proves to be far from unique.

When searching for 'Historische Steekproef Nederlandse bevolking', all hits refer to the HSN, but all of them indirectly. The first is to historici.nl, which links on to the website of the International Institute of Social History, and several among the first ten hits are to that institute's website. But once there, it takes some navigational skills and at least four clicks, before we actually reach the HSN. Like with TRA, the result of searching with 'database historical demography Netherland' is much better, delivering a link to the HSN in the first hit. Again, just like TRA, this is through https://ehps-net.eu. And again, just like TRA, HSN is an abbreviation of lots of other things besides 'our' HSN, from the *Horeca Stichting Nederland* to the *Hervormde Schoolvereniging te Nijkerk*. Like Umeå, HSN offers a large number of data sets which have their own name. Incidentally, one of these is an acronym also used in Umeå: CEDAR, in this case standing for Census Data Open Linked. HSN uses these acronyms typically to distinguish between data sets that have been added to the database by different researchers in differently funded projects. But the website makes clear what the HSN is, and how the different subsets contribute to it.

When searching for IPUMS, the first hit directs to IPUMS' home site. However, searching for 'database historical demography US' gives mainly hits that direct to the US census. The third hit relates to the International Committee for Historical Demography and the seventh hit is the Library at Berkeley. Both mention IPUMS among other, similar databases. IPUMS' home page lists a number of different data sets, which all are named IPUMS, plus a word or acronym to make clear in which way the specific data sets differ, like IPUMS-USA, IPUMS-International, IPUMS-Global Health or IPUMS-NHGIS (National Historical GIS).

Before we conclude on findability, it is important to stress that all of these databases make dispersed and hard to access data extremely more findable than they would be otherwise. What we have looked at here, is merely how findable the resulting databases are. One could well argue that that is just scratching the surface of findability. But that said, a few clear recommendations are possible:

- The best way to make your database findable is by using a distinctive acronym or name. IPUMS is much better than TRA or HSN in that way. Being known by the same name as your town or university does not increase findability.
- Only being findable by the exact correct way to spell your name is not conducive to being found.
- Having lots of references to your original website with different search terms enhances findability. This is mostly the ordinary management of web site findability, but international forms of association of these databases or crosslinking can be tremendously helpful (see note 1).
- Stick to your name. Obviously, this may run counter to some of the other recommendations if you have chosen the wrong name originally.
- Name your data sets after the main database, and find other ways to acknowledge contributions from researchers and funders.

3 ACCESSIBLE

How *accessible* are the data sets? Individual demographic data are sensitive. Some of the data sets contain medical information, like cause of death or the results of a medical examination for military service. Having spent time in an institution or having lived in certain neighborhoods can also be deemed sensitive, as well as belonging to an ethnic group or religious community. Historians generally do not think that the deceased have a right to privacy, but some of this information can be sensitive across generations. All four databases are careful about the way sensitive information may spread, and want their users to adhere to relevant scholarly standards.

Both TRA and HSN only deliver data after one has requested and acquired permission from a database representative, whom can be contacted through e-mail. HSN users must fill in and sign a license agreement and agree with HSN privacy rules. TRA lists eight conditions that users must underwrite, and is therefore slightly more transparent than HSN. Both Umeå and IPUMS restrict the use of data that may contain information on living individuals. Umeå asks users to specify which data will be retrieved and how they will be used. This is reviewed by the database's Approval Committee, and in case of sensitive personal data also by the Swedish Ethical Review Authority. IPUMS-International uses what is describes as a 'lengthy, probing registration form' which it deems 'an effective deterrent for unqualified applicants'.

I did not bother the officials at TRA and HSN by requesting access just to experience the process. Both IPUMS and Umeå make available data sets which are anonymized or which contain no data on surviving individuals. Umeå makes available anonymous data for the older period without further ado through its application SHiPS. IPUMS asked me to check two boxes promising not to redistribute the data without permission, to cite the data appropriately and to add publications making use of IPUMS data to their bibliography (Ruggles et al., 2020). Similar wishes are expressed by all databases. Giving my email address and answering a small number of questions resulted in an immediate response from IPUMS granting me access. The application which allowed me to download data from IPUMS was user-friendly, SHiPS somewhat less so.

Recommendations again are possible:

- Be transparent about the conditions you set for using data.
- Let users of privacy-sensitive data describe how they will use the data and ask them to state that they will comply with your rules of good scholarly practice. Let their application be judged by the database and — if necessary — by an ethical review board.
- Make non-sensitive data available upon simple request and grant access to them through a quick and automatic procedure.

4 INTEROPERABLE

Are the databases *interoperable*? The proof of the pudding would be in linking relevant data sets of our four databases. That is beyond the scope of this contribution, as the fact that only two data sets were accessed already makes clear. Fortunately, others have already tasted the pudding.

On theoretical grounds alone we would assume that data sets based on census data and covering recent decades would stand a better chance to collect the same data defined in a similar way, than data sets based on register data and covering an older period. From 2006 an Intermediate Data Structure (IDS) was developed to improve interoperability among the databases (Alter & Mandemakers, 2014; Quaranta, 2015). This makes it possible to download data from two of the databases into the IDS and compare them. Using this tool, it was for instance possible to link the HSN and Umeå databases, and two others for Scania and Antwerp, and help establish that infant mortality runs in the family: the likelihood that a woman's offspring died in infancy was higher when any of her siblings had died in infancy (Quaranta & Sommerseth, 2018).

It is also possible to follow individuals from one database to the other. Paiva, Anguita and Mandemakers (2020) did this when they followed Dutch migrating to the USA by linking HSN to IPUMS. This exercise points to a difference between census-based and register-based data sets. Both have to link individuals and households that appear at different places in the original sources. For the census data, linking them from one observation to the next means bridging the time lapse between two censuses, typically ten years. A child's whole life course can fall between two censuses and thus will not be visible in any census. Registers follow individuals and households through time. Apart from additional richness in observed geographical movements and life events, this also makes it easier to link people from one register to another than from one census to the next.

Dataverse versions of the HSN data (https://datasets.iisg.amsterdam/dataverse/hsn) come with help to enhance interoperability: a suite of tools for the classification and comparison of historical occupational titles (HISCO, HISCAM and HISCLASS), and a gazetteer of place names in the Netherlands in the relevant period (AMCO). IPUMS used an adapted version of the HISCO classification for several of its data sets (IPUMS USA 1880 100% population database, IPUMS NAPP = North Atlantic Population Project). A conversion table between HISCO and the US Bureau of the Census 1950 Standard for Occupations, CROSSWALK, is available (https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:73810).

5 RESUSABLE

Being *reusable* is clearly a strong point of all these data sets. The investment necessary to establish the databases in the first place is far too large to be justified by answering a single historical research question. They are specifically designed to be used over and over again. They often grow when they are reused and the additional data needed to answer a fresh research question are entered in the same format and added to the database.

The four databases we discussed all are fair, but some are fairer than others. The conclusions are summarized in the table below.

Table 1 *FAIRness of four on line databases on historical life courses*

	TRA	Umeå	HSN	IPUMS
Findable online				
with acronym/name the database is generally known by	*	*****	*	*****
with 'database historical demography <country>'	***	***	***	*
use of other acronyms to designate separate data sets	****	*	***	*****
Accessible	**	****	*	*****
Interoperable	****	****	****	*****
richness of data	*****	*****	*****	***
linkability on individual level	****	****	*****	**
Reusable	*****	*****	*****	*****

Note: The number of stars indicates the degree of compliance with the element of FAIRness, with 5 stars as maximum.

Interoperability and reusability are part of the DNA of these databases, and it should come as no surprise that they score well on these rows. Limitations to interoperability are caused by characteristics of the historical sources, like censuses providing data windows with ten year intervals. Where interoperability is limited by differences in the ways databases have processed data, tools like the IDS come to the rescue. Not all databases score as well on being Findable and Accessible, as defined here. Relatively simple measures could raise the Findability and Accessibility of the databases that have lower scores on these rows.

REFERENCES

Alter, G., & Mandemakers, K. (2014). The Intermediate Data Structure (IDS) for longitudinal historical microdata, version 4. *Historical Life Course Studies*, *1* 1–26. Retrieved from http://hdl.handle.net/10622/23526343-2014-0001?locatt=view:master

Bourdieu, J., Kesztenbaum, L., Postel-Vinay, G. (2014). L'enquête TRA, histoire d'un outil, outil pour l'histoire: Tome 1 (1793–1902). Paris, INED: Classiques de l'économie et de la population.

Edvinsson, S. & Engberg, E. (2020). A database for the future. Major contributions from 47 years of database development and research at the Demographic Data Base. *Historical Life Course Studies, Online first.* Retrieved from https://hdl.handle.net/10622/23526343-2020-0009

Mandemakers, K. & Kok, J. (2020). Dutch Lives. The Historical Sample of the Netherlands (1987–): Development and research. *Historical Life Course Studies, Online first*. Retrieved from http://hdl.handle.net/10622/23526343-2020-0001?locatt=view:master

Paiva, D., Anguita, F., & Mandemakers, K. (2020). Linking the Historical Sample of the Netherlands with the USA censuses, 1850–1940. *Historical Life Course Studies, 9*, 1–23. Retrieved from http://hdl.handle.net/10622/23526343-2020-0003?locatt=view:master

Quaranta, L. (2015). Using the Intermediate Data Structure (IDS) to construct files for statistical analysis. *Historical Life Course Studies, 2*, 86–107. Retrieved from http://hdl.handle.net/10622/23526343-2015-0007?locatt=view:master

Quaranta, L., & Sommerseth, H. L. (2018). Introduction: Intergenerational transmissions of infant mortality using the Intermediate Data Structure (IDS). *Historical Life Course Studies, 7,* 1–10. Retrieved from http://hdl.handle.net/10622/23526343-2018-0014?locatt=view:master

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J., & Sobek, M. (2020). *IPUMS USA: Version 10.0* [Data set]. Minneapolis, MN: IPUMS. doi: 10.18128/D010.V10.0