# Reflections on the Use of the Intermediate Data Structure (IDS) in Historical Demographic Research

By Luciana Quaranta

## 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

# Reflections on the Use of the Intermediate Data Structure (IDS) in Historical Demographic Research

Luciana Quaranta

Lund University

## ABSTRACT

The Intermediate Data Structure (IDS) was developed as a strategy aimed at standardizing the dissemination of micro-level historical demographic data. The structure provides a common and clear data strategy which facilitates studies that consider several databases, and the development and exchange of software. Based on my own experiences from working with the IDS, in this article I provide reflections on the use of IDS to create datasets for analysis and to conduct comparative demographic research.

**Keywords:** Intermediate Data Structure, Historical demography, Comparative research, Longitudinal data

# 1    INTRODUCTION

The Intermediate Data Structure (IDS) was developed by George Alter, Myron Gutmann and Kees Mandemakers as a strategy aimed at standardizing the dissemination of micro-level historical demographic data (Alter & Mandemakers, 2014; Alter, Mandemakers, & Gutmann, 2009). The structure provides a common and clear data strategy which facilitates conducting comparative studies using multiple databases and the development and exchange of software for analysis, therefore increasing the transparency, replicability and generalizability of research and breaking down the barriers of entry into the field of historical demography. Using longitudinal demographic data is, in fact, complex, given its multilevel and relational character and the fact that it involves processes that develop over time.

The article by George Alter (2021) included in the current issue examines the main characteristics of IDS and how these characteristics make the structure flexible and expandable. In the current work I provide reflections on the use of IDS to create data sets for analysis and to conduct comparative demographic research. These reflections are based on my own experiences of working with the IDS of the Scanian Economic Demographic Database and from the development of an international comparative project that used IDS databases from five different historical European populations.

# 2    REFLECTIONS ON THE USE OF IDS TO CREATE DATA SETS FOR ANALYSIS

Using the IDS requires initial investments. Foremost, it is necessary to transfer the data from its original form into the standardized structure. This can be done either using your own-written programs or the publicly available IDS transposer (Klancher Merchant & Alter, 2017). There are large returns to the investments made into transferring the data into IDS, since it provides a clear and well-defined structure and, through the METADATA table, good documentation of variables and their values, all of which substantial facilitate the work of researchers. The common structure makes it also easier for researchers to switch from using one database to another, or to conduct comparative studies. Moreover, the possibility of using publicly available software (e.g. Alter, Newton, & Oeppen, 2020; Quaranta, 2016, 2018b) reduces the barriers of entry into the field for researchers who are less experienced in data management.

To conduct longitudinal statistical analyses using data stored in the IDS it is necessary to select from the IDS tables the information that is required for the study, to process such data, to construct additional variables, and to convert the data extraction into a rectangular episodes table. This process is not simple. The IDS follows the entity-attribute-value model, and therefore contains one row for each attribute (variable) about an entity (person, place, etc.). Although this model has the advantage of making the structure clear and standardized, extracting variables from the tables can be somewhat tedious. For variables that are used multiple times in the process of creating data sets for analysis (birth date, sex, etc.), it can sometimes be useful to construct intermediate wide tables. For other variables, even if extracting them from the IDS can require some extra lines of code with respect to using more traditional database structures, due to the standard nature of IDS, and with simple modifications the same code can be used to process different variables, for example through the use of loops.

In a previous article (Quaranta, 2015), I proposed a solution for creating rectangular data sets for analysis from IDS databases, which involved the use of the Chronicle file. The Chronicle file is a middle step between the IDS tables — which store the information found in sources — and the episodes tables — which are rectangular files that are ready for statistical analysis. Like the basic IDS tables, the Chronicle file follows the entity-attribute-value model, and it contains one row per each date of an event and each date of change in variable values. Using the program 'Episodes File Creator' (Quaranta, 2016), the Chronicle file can be transformed into a rectangular table for statistical analysis. One advantage of using the Chronicle file when developing a data set for analysis from IDS-data is that it allows to create variables in a modular way. In other words, rather than creating all variables at the same time, it is possible to develop separate syntax or sections of syntax files, each focusing on a specific variable (marital status, occupation, household size, etc.) or event (marriage, birth of a child, death, etc.). Programming variables in a modular fashion does not increase computational time, and it allows to easily check, modify and run variable-specific code, and to use the same constructed variables in different types of analyses, in the same way as source variables stored in the basic IDS tables. Modular programming also facilitates the exchange of software between researchers, since it is possible not only to share software that produces full data sets for analysis but also to share software that produces only one specific variable.

In 2009, together with research engineer Clas Andersson, I have converted the Scanian Economic Demographic Database into the IDS (see Dribe and Quaranta (2020) for a full description on the SEDD source material, structure of the data and IDS transfer). In addition to creating basic IDS tables containing source information, using such source information we constructed a wide range of individual- and context-level variables, to be used for many different types of research questions (related to fertility, marriage, migration, social mobility, mortality, etc.). All these variables are included in a Chronicle file. Researchers who request data from SEDD are provided with both basic IDS tables and a Chronicle file, which allows them to conduct most types of analyses without further elaboration of the data. This has not only facilitated the work of researchers, but it has also made different research outputs more consistent, since variables were defined identically. The modular nature of the Chronicle file and of the programs used to create variables has allowed to, when necessary, easily update and fix the code specific to each constructed variable as new data was added to SEDD. It has also allowed to add new modules to create additional variables when research questions were expanded. From these experiences in working with IDS in SEDD I can therefore say that IDS functions well, it is very advantageous for research, and that even if the investments made are large, the returns to these investments are much larger.

## 3    REFLECTIONS ON USING IDS IN COMPARATIVE RESEARCH

The growing availability of micro-level longitudinal databases combined with the development of advanced statistical techniques and related software has allowed major developments in the field of historical demography. However, due to the high costs involved in the digitization of historical sources, the majority of the research that has been published has been based on studies that focus on small communities and on limited time periods, making it difficult to test the external validity of their findings.

One of the best ways to extend the scope of small databases and the generalizability of results across different contexts and time periods is by conducting comparative research that considers several populations. Facilitating comparative research was one of the main aims in the development of the IDS. The first and to date also the only international comparative research project based on the IDS looked into intergenerational transmissions in infant mortality across the maternal line in five different historical populations in Europe (Quaranta & Sommerseth, 2018) — Belgium (Donrovich, Puschmann, & Matthijs, 2018), the Netherlands (van Dijk & Mandemakers, 2018), Norway (Sommerseth, 2018), northern Sweden (Broström, Edvinsson, & Engberg, 2018) and southern Sweden (Quaranta, 2018a). The studies were conducted using the same methods and theoretical framework, and the same programs were used to create a Chronicle file for each database (Quaranta, 2018b), transform such file into rectangular a episodes table (Quaranta, 2016) and to run the statistical models (Quaranta, 2018b).

Developing the software used in the project to create the Chronicle file was complex. However, such complexity was related to the underlying characteristics of the research question and data, which required following individuals longitudinally across their lives and across generations, rather than the fact that five different databases were used. Writing one common program to be used across the five databases included in the project was not much more difficult nor much more time consuming than writing a program to be used by one single database. The only additional challenge faced in developing a common syntax code was in the definition of when individuals were under exposure, given that some of the databases contained only information on vital events, while other databases included data on vital events as well as data on migration.

Thanks to the IDS, being able to use the same exact programs in our study was a novelty with respect to previous works. The few comparative historical demographic studies that had been published before were primarily based on common research questions and model definitions across the communities considered, but, given the different underlying structure of the databases used, they employed independent syntax for each database to create the data sets for analysis and run the statistical models. Such studies were unable to distinguish differences in the findings that related from diverging characteristics of the contexts from those related to differences in data handling and model generating processes. Instead, in our IDS-based study, it was possible to fully compare the findings and even to pool the data sets to estimate one common model across the five populations. The project demonstrated that the IDS and the construction of Chronicle files do indeed greatly facilitate the development of collaborative research and allow to make studies fully comparative and reproducible. Given that the software used in the project was published, the same software could be used also

in other contexts. The project also demonstrated that this approach can be utilized in many new IDS-based comparative projects, thereby expanding the future of historical demographic research.

# 4 CONCLUSIONS

The IDS was developed with the aim of providing a standardized structure to facilitate the dissemination of micro-level historical demographic data. My own experience with the use of the IDS has shown that, although the initial investments needed to transfer the data into IDS and to learn how to use the structure are not negligible, there are very large returns to these investments. Detailed documentation is provided to researchers. Data sets for analysis can be developed by first creating Chronicle files, and in a modular way, simplifying the control, expansion and exchange of programs. The standard nature of IDS also greatly facilitates the development of comparative studies. The IDS therefore has the potential of breaking the barriers of entry into the field of historical demography, of making research more transparent and reproducible, and results generalizable to wider contexts, in these ways expanding the scope of future historical demographic research.

To date not many standard IDS programs have been published, something perhaps due to the fact that documenting software is time consuming and that software development is not yet recognized as academic merit within our discipline. The organization of research workshops devoted to writing IDS programs can be a step-forward in this development. The establishment of more research collaborations involving IDS-based comparative studies can also help to incentivise more databases to transfer their data into the IDS and researchers to produce and share IDS-software. The main Swedish historical demographic databases, including SEDD, have joined in the national infrastructure SwedPop (www.swedpop.se), which will allow researchers to download IDS-based harmonized data from the databases. This infrastructure will not only significantly improve access to Swedish population data, but it will also enable more comparative studies to be conducted and possibly also furtherly promote the use of IDS internationally.

# REFERENCES

Alter, G. (2021). Reflections on the Intermediate Data Structure (IDS). *Historical Life Course Studies, 10*, 71–75. doi: 10.51964/hlcs9570

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

Alter, G., Mandemakers, K., & Gutmann, M. (2009). Defining and distributing longitudinal historical data in a general way through an intermediate structure. *Historical Social Research, 34*(3), 78–114. doi: 10.12759/hsr.34.2009.3.78-114

Alter, G., Newton, G., & Oeppen, J. (2020). Re-introducing the Cambridge Group Family Reconstitutions. *Historical Life Course Studies*, *9*, 24–48. Retrieved from https://hdl.handle.net/10622/23526343-2020-0005

Broström, G., Edvinsson, S., & Engberg, E. (2018). Intergenerational transfers of infant mortality in 19th-century northern Sweden. *Historical Life Course Studies*, *7*, 106–122. Retrieved from http://hdl.handle.net/10622/23526343-2018-0005?locatt=view:master

Donrovich, R., Puschmann, P., & Matthijs, K. (2018). Mortality clustering in the family. Fast life history trajectories and the intergenerational transfer of infant death in late 19th- and early 20th-century Antwerp, Belgium. *Historical Life Course Studies*, *7*, 47–68. Retrieved from http://hdl.handle.net/10622/23526343-2018-0006?locatt=view:master

Dribe, M., & Quaranta, L. (2020). The Scanian Economic-Demographic Database (SEDD). *Historical Life Course Studies*, *Online first*. Retrieved from https://hdl.handle.net/10622/23526343-2020-0008

Klancher Merchant, E., & Alter, G. (2017). IDS Transposer: A users guide. *Historical Life Course Studies*, *4*, 59–96. Retrieved from http://hdl.handle.net/10622/23526343-2017-0004?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. (2016). STATA programs for using the intermediate data structure (IDS) to construct files for statistical analysis. *Historical Life Course Studies, 3*, 1–19. Retrieved from http://hdl.handle.net/10622/23526343-2016-0001?locatt=view:master

Quaranta, L. (2018a). Intergenerational transfers in infant mortality in southern Sweden, 1740–1968. *Historical Life Course Studies, 7*, 88–105. Retrieved from http://hdl.handle.net/10622/23526343-2018-0013?locatt=view:master

Quaranta, L. (2018b). Program for studying intergenerational transmissions in infant mortality using the Intermediate Data Structure (IDS). *Historical Life Course Studies*, *7*, 11–27. Retrieved from http://hdl.handle.net/10622/23526343-2018-0010?locatt=view:master

Quaranta, L., & Sommerseth, H. (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

Sommerseth, H. (2018). The intergenerational transfer of infant mortality in northern Norway during the 19th and early 20th centuries. *Historical Life Course Studies*, *7*, 69–87. Retrieved from http://hdl.handle.net/10622/23526343-2018-0008?locatt=view:master

van Dijk, I. K., & Mandemakers, K. (2018). Like mother, like daughter. Intergenerational transmission of infant mortality clustering in Zeeland, the Netherlands, 1833–1912. *Historical Life Course Studies, 7*, 28–46. Retrieved from http://hdl.handle.net/10622/23526343-2018-0003?locatt=view:master