Using the Intermediate Data Structure (IDS) to Construct Files for Statistical Analysis

Quaranta, Luciana

Published in:
Historical Life Course Studies

2015

HISTORICAL LIFE COURSE STUDIES

VOLUME 2
2015

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 EHPS-Net 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 http://www.ehps-net.eu/journal.

Editors: Koen Matthijs & Paul Puschmann
Family and Population Studies
KU Leuven, Belgium
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.

EUROPEAN SCIENCE FOUNDATION

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.

EHPS NETWORK

HISTORICAL LIFE COURSE STUDIES
VOLUME 2 (2015), 86-107 published 18-12-2015

Using the Intermediate Data Structure (IDS) to Construct Files for Statistical Analysis

Luciana Quaranta
Lund University

ABSTRACT

The use of longitudinal historical micro-level demographic data for research presents many challenges. The Intermediate Data Structure (IDS) was developed to try to solve some of these challenges by facilitating the storing and sharing of such data. This article proposes an extension to the IDS, which allows the standardization and storage of constructed variables. It also describes how to produce a rectangular episodes file for statistical analysis from data stored in the IDS and presents programs developed for such purpose.

Keywords: Intermediate Data Structure, IDS, Historical Demography, Demography, Life Courses, Social History, History, Episodes Table, Survival Analysis, Event History Analysis, STATA

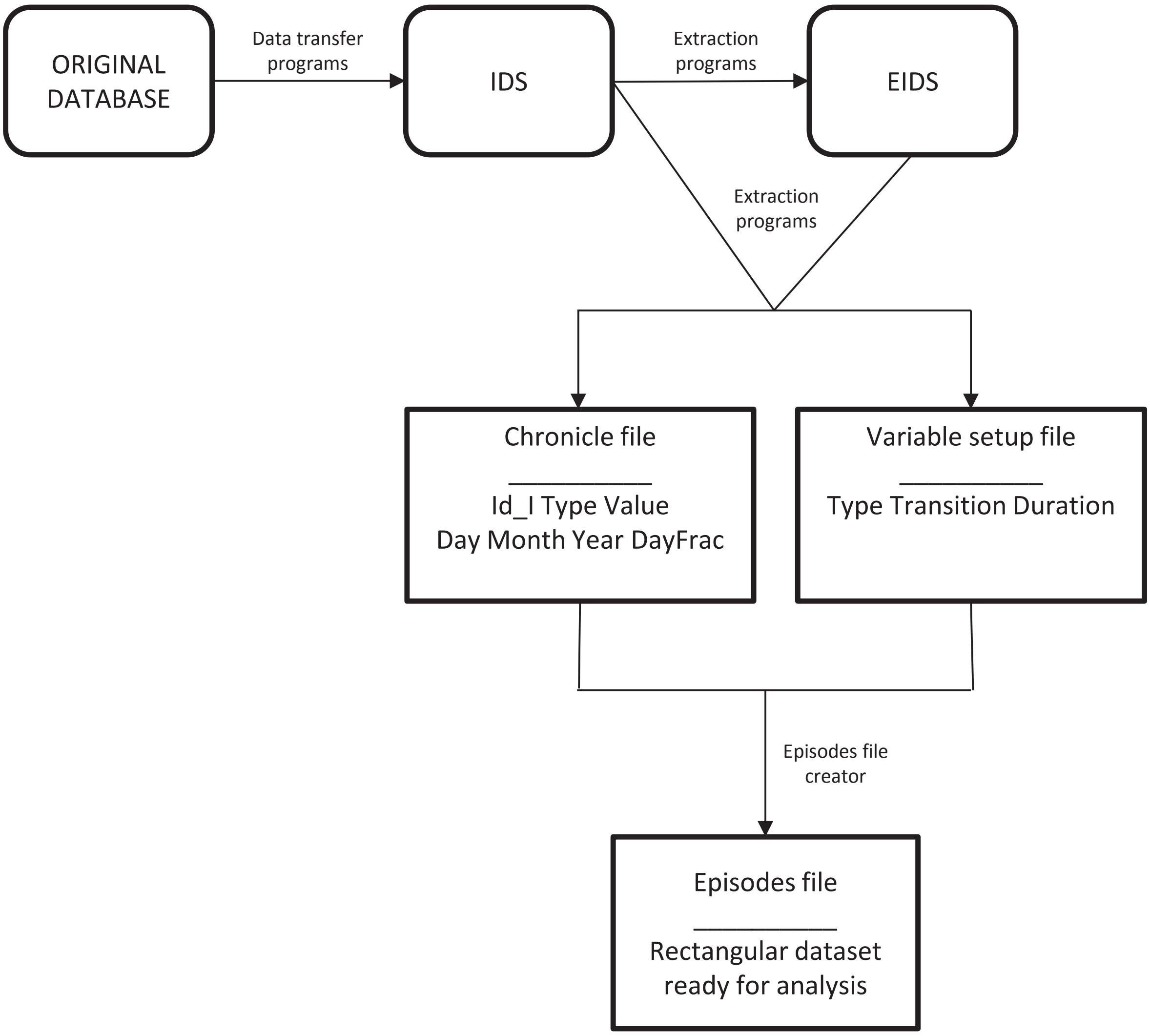
e-ISSN: 2352-6343
PID article: http://hdl.handle.net/10622/23526343-2015-0007?locatt=view:master

1 INTRODUCTION

Over the past decades the field of historical demography has greatly expanded and many important findings have been discovered. This is partly due to the availability of digitized historical longitudinal micro-level databases such as the Scanian Economic Demographic Database (Bengtsson et al. 2014), the Historical Sample of the Netherlands (Mandemakers 2000), and the Demographic Database at Umeå University (Danell 1981; Edvinsson 2000). These types of databases have allowed us to improve our understanding about, among other things, mortality (e.g. Bengtsson & van Poppel 2011; Schofield, Reher & Bideau 1991), fertility (e.g. Bengtsson & Dribe 2014; Quaranta 2011; Reher & Sanz-Gimeno 2007; Van Bavel 2004), social stratification and social mobility (e.g. Dribe, Van Bavel & Campbell 2012), the long-term impacts of early life conditions (e.g. Bengtsson & Lindström 2003; Lindeboom, Portrait & van den Berg 2010; Öberg 2014; Quaranta 2013, 2014) and the short-term impact of economic stress (e.g. Bengtsson, Campbell & Lee 2004; Lundh & Kurosu 2014; Tsuya et al. 2010).

The use of longitudinal micro-level historical demographic data presents many challenges which are often connected to their multilevel and relational aspects, as well as to the complexity of conceptualizing processes that develop over time. The data allow researchers to conduct studies that test hypotheses relating to sequential events over life histories, for example through the use of event history analysis techniques such as Cox models (Cox 1972; Therneau & Grambsch 2000). In order to conduct research using event history analysis, the data have to be set up as rectangular episodes files. Creating such files requires advanced data management skills. Although today there are numerous kinds of statistical software available to conduct complex statistical modeling, there are not many programs that perform automated data management and that can set up longitudinal micro-level demographic data

Figure 1 *Schema of the transformation of data from the original database to a rectangular episodes file*



in the correct format for analysis. The lack of programs and structures restricts the usability of such data to researchers with extensive programming skills, and therefore also limits the scope of historical demographic research.

The Intermediate Data Structure (IDS) was developed as a strategy aimed at simplifying the collecting, storing and sharing of historical demographic data (Alter & Mandemakers 2014; Alter, Mandemakers & Gutmann 2009). The structure provides a common platform to store data from different databases, regardless of their original form. Among other advantages, the structure facilitates the sharing of data and software and increases the transparency of how data outputs are prepared. Previous articles have discussed in detail how to store data in the IDS, but not much has been written about how to extract data out of the IDS to produce files for research.

To conduct longitudinal statistical analyses using data stored in the IDS it is necessary to select the information that is required for the study from the IDS tables in order to process such data for the construction of additional variables and to convert the data extraction into a rectangular episodes table. This article presents different concepts for creating an episodes table for statistical analysis from data stored in the IDS. It discusses a series of steps for creating such files, which are summarized in Figure 1. This work also introduces seven open-access programs which can be used to conduct such steps. A detailed explanation of the structure and use of these programs is published in Quaranta (Forthcoming), with the programs also attached.

While presenting the aforementioned steps and programs, this article also proposes an extension of the IDS, the Extended Intermediate Data Structure (EIDS). The IDS is intended for the storage of data obtained directly from the sources (e.g. date of birth, individual occupation, date of marriage, address). The principle behind the EIDS is that the same structure of tables can also be used to store constructed variables (e.g. household size, civil status, household head occupation). Such an extension allows researchers and database administrators to reutilize and share constructed variables which expands the range of possible users of a database as well as the transparency and replicability of research studies. In addition, this article proposes a format that data extractions should follow in order to facilitate the conversion of this data into rectangular episodes files for analysis. The output given by extraction programs developed to be used with the IDS should also follow this format in order to be able to combine data produced by different extraction programs.

This work is developed within the initiative of the European Historical Samples Network (EHPS-net) which, among other things, has the aim of creating extraction software to store data into the IDS and programs to use such stored data for analysis. The article is targeted towards researchers, programmers and database administrators. The solutions and programs introduced can be applied to historical demographic databases created from population registers or family reconstitutions. These solutions and programs can be used to extract a dataset for analysis linked to any type of research question dealing with longitudinal analysis, as long as the data has been transformed into the IDS.

The text is organized into different sections. After briefly describing the tables included in the IDS, the EIDS is introduced and its advantages are discussed. A description of how to store data in the EIDS is given next. Finally, the steps and programs that can be used to select data stored in the IDS and the EIDS and to build a file for analysis from such extractions are presented.

2 EXTENDING THE IDS TO INCLUDE CONSTRUCTED VARIABLES

The IDS consists of five main tables: INDIVIDUAL which is used to store information relating to individuals; INDIV_INDIV which defines relationships between individuals; CONTEXT which defines geographical contexts and contains information about them; CONTEXT_CONTEXT which defines contexts that are nested in other higher level contexts; INDIV_CONTEXT which defines spells of times during which individuals are present in a specific context. The IDS also contains a METADATA table which is used to describe the variable Types and Values included in the five main tables. Data stored in the IDS follows the Entity Attribute Data Model (Stead, Hammond & Straube 1982), and therefore contains one attribute per each record. In other words, each row of the table contains one declaration of a Value of a variable Type. A detailed description of how to store data into the IDS is given in Alter & Mandemakers (2014).

Data can be stored in the IDS using data transfer programs, as was shown in Figure 1. Such programs can be created locally, or publicly available programs can also be used. Members of the EHPS-net are currently creating an open-access web system aimed at transforming data into the IDS.

The data stored in the IDS can be used in different types of analyses. The purposes and definition of each study determine which information and variables must be used. Some of these variables consist of information that is available in the sources and can be obtained directly from the IDS tables, while other variables need to be constructed from calculations and/or elaborations of data stored in the IDS tables. Open-access extraction programs made available for the IDS community or locally created functions can be used to select and create variables.

Tables 1-5 show examples of the five main IDS tables. The information contained in these tables was created to resemble data that would be obtained from population registers. This made-up data is used throughout this article to present the EIDS and the seven STATA programs that were developed to select and elaborate such data.

Table 1 *Example of an INDIVIDUAL table*

Id_D	Id_I	Type	Value	Value_Id_C	Day	Month	Year	Date_type	Source
Test_DB	1148964	Birth_date			8	11	1804	Declared	Marriage_register
Test_DB	1148964	Marriage			11	4	1825	Event	Marriage_register
Test_DB	1148964	Occupation	Farmer		11	4	1825	Declared	Marriage_register
Test_DB	1148964	Start_observation	Arrival		11	4	1825	Declared	Parish_register
Test_DB	1148964	Death			8	11	1855	Event	Death_register
Test_DB	1148964	End_observation	Death		8	11	1855	Declared	Death_register
Test_DB	1148964	Sex	Male					Declared	Marriage_register
Test_DB	1148964	Birth_location		12478				Declared	Birth_register
Test_DB	1237852	Birth_date			12	4	1807	Declared	Birth_register
Test_DB	1237852	Marriage			11	4	1825	Event	Marriage_register
Test_DB	1237852	Start_observation	Arrival		11	4	1825	Declared	Parish_register
Test_DB	1237852	Death			15	4	1838	Event	Death_register
Test_DB	1237852	End_observation	Death		15	4	1838	Declared	Death_register
Test_DB	1237852	Sex	Female					Declared	Marriage_register
Test_DB	1237852	Birth_location		14789				Declared	Birth_register
Test_DB	1378563	Birth			18	2	1853	Event	Birth_register
Test_DB	1378563	Birth_date			18	2	1853	Declared	Birth_register
Test_DB	1378563	Start_observation	Birth		18	2	1853	Declared	Birth_register
Test_DB	1378563	End_observation	Departure		1	12	1874	Declared	Parish_register
Test_DB	1378563	Sex	Female					Declared	Birth_register
Test_DB	1378563	Birth_location		11111				Declared	Birth_register
Test_DB	1479856	Birth_date			15	12	1831	Declared	Birth_register
Test_DB	1479856	Start_observation	Arrival		1	2	1852	Declared	Parish_register
Test_DB	1479856	End_observation	Departure		14	11	1881	Declared	Parish_register
Test_DB	1479856	Birth_location		11111				Declared	Birth_register
Test_DB	1479856	Sex	Male					Declared	Birth_register
Test_DB	1548468	Birth			7	10	1855	Event	Birth_register
Test_DB	1548468	Birth_date			7	10	1855	Declared	Birth_register
Test_DB	1548468	Death			7	10	1855	Event	Death_register
Test_DB	1548468	End_observation	Death		7	10	1855	Declared	Death_register
Test_DB	1548468	Start_observation	Birth		7	10	1855	Declared	Birth_register
Test_DB	1548468	Birth_location		11111				Declared	Birth_register
Test_DB	1548468	Sex	Female					Declared	Birth_register

Table 1 continued on next page

Id_D	Id_I	Type	Value	Value_Id_C	Day	Month	Year	Date_type	Source
Test_DB	1567526	Birth_date			26	7	1821	Declared	Marriage_register
Test_DB	1567526	Marriage			16	9	1851	Event	Marriage_register
Test_DB	1567526	Start_observation	Arrival		16	9	1851	Declared	Parish_register
Test_DB	1567526	Death			4	8	1885	Event	Death_register
Test_DB	1567526	End_observation	Death		4	8	1885	Declared	Death_register
Test_DB	1567526	Sex	Female					Declared	Marriage_register
Test_DB	1567526	Birth_location		12478				Declared	Birth_register
Test_DB	1897563	Birth_date			5	8	1819	Declared	Birth_register
Test_DB	1897563	Start_observation	Arrival		17	11	1836	Declared	Parish_register
Test_DB	1897563	End_observation	Departure		12	1	1852	Declared	Parish_register
Test_DB	1897563	Birth_location		22222				Declared	Birth_register
Test_DB	1897563	Sex	Male					Declared	Birth_register
Test_DB	1945568	Birth			18	11	1828	Event	Birth_register
Test_DB	1945568	Birth_date			18	11	1828	Declared	Birth_register
Test_DB	1945568	Start_observation	Birth		18	11	1828	Declared	Birth_register
Test_DB	1945568	Marriage			16	9	1851	Event	Marriage_register
Test_DB	1945568	Occupation	Farmhand		16	9	1851	Declared	Marriage_register
Test_DB	1945568	Occupation	Farmer		18	2	1853	Declared	Birth_register
Test_DB	1945568	Death			3	6	1878	Event	Death_register
Test_DB	1945568	End_observation	Death		3	6	1878	Declared	Death_register
Test_DB	1945568	Sex	Male					Declared	Birth_register
Test_DB	1945568	Birth_location		22222				Declared	Birth_register

Table 2 *Example of a CONTEXT table*

Id_D	Id_C	Type	Value
Test_DB	11111	Level	Parish
Test_DB	11111	Name	Lund
Test_DB	12345	Level	Household
Test_DB	12478	Level	Parish
Test_DB	12478	Name	Kävlinge
Test_DB	14789	Level	Parish
Test_DB	14789	Name	Hög
Test_DB	22222	Level	Parish
Test_DB	22222	Name	Malmö
Test_DB	35891	Level	Household

Table 3 *Example of an INDIV_CONTEXT table*

Id_D	Id_I	Id_C	Start_day	Start_month	Start_year	End_day	End_month	End_year
Test_DB	1148964	12345	11	4	1825	8	11	1855
Test_DB	1237852	12345	11	4	1825	15	9	1836
Test_DB	1945568	12345	18	11	1828	8	6	1849
Test_DB	1897563	12345	17	11	1836	12	1	1852
Test_DB	1237852	12345	10	8	1837	15	4	1838
Test_DB	1567526	35891	16	9	1851	4	8	1885
Test_DB	1945568	35891	16	9	1851	3	6	1878
Test_DB	1479856	35891	1	2	1852	14	11	1881
Test_DB	1378563	35891	18	2	1853	1	12	1874
Test_DB	1548468	35891	7	10	1855	7	10	1855

Table 4 *Example of a CONTEXT_CONTEXT table*

Source	Id_C_1	Id_C_2	Relation
Test_DB	12345	12478	Household and parish
Test_DB	35891	14789	Household and parish

Table 5 *Example of an INDIV_INDIV table*

Id_D	Id_I_1	Id_I_2	Relation
Test_DB	1148964	1945568	Child
Test_DB	1148964	1897563	Servant
Test_DB	1148964	1237852	Wife
Test_DB	1237852	1945568	Child
Test_DB	1237852	1148964	Husband
Test_DB	1378563	1945568	Father
Test_DB	1378563	1567526	Mother
Test_DB	1479856	1945568	Master
Test_DB	1548468	1945568	Father
Test_DB	1548468	1567526	Mother
Test_DB	1567526	1378563	Child
Test_DB	1567526	1548468	Child
Test_DB	1567526	1945568	Husband
Test_DB	1897563	1148964	Master
Test_DB	1945568	1378563	Child
Test_DB	1945568	1548468	Child
Test_DB	1945568	1148964	Father
Test_DB	1945568	1237852	Mother
Test_DB	1945568	1479856	Servant
Test_DB	1945568	1567526	Wife

The aim of the EIDS is to store variables that are constructed from source information in the IDS. Constructed variable types can be referred to as extended variables. The EIDS has the same structure as the IDS and is based on the tables INDIVIDUAL_EXT and CONTEXT_EXT. The METADATA table of the IDS can be expanded to also include information on the definition of variables and variable values of data stored in the EIDS.

The use of the EIDS to store extended variables has various advantages. Extended variables are often used repeatedly for different types of analyses and research projects. By storing such variables in the EIDS they can be reused without a need for their reconstruction. Moreover, the treatment of information and the construction of extended variables for analysis often require a series of decisions to be taken by database managers or experts of the sources of a specific database. By storing these variables in the EIDS such decisions only need to be taken once. Accordingly, all users of a certain database consider the same definition of their variables which reduces the possibility of discrepancies between studies and increases replicability and transparency. Detailed descriptions of variable definitions can be included in the METADATA table. In addition, research projects that focus on the same database can share extended variables.

The use of the EIDS can also expand the range of possible users of a database. It allows students and other less experienced researchers to use this data in their analysis without having to conduct complex programming. The EIDS also reduces the time, computational power and resources needed to obtain and produce a dataset for analysis. Such datasets could be created by simply using an extraction program to select from a list of available variables, and an episodes file creator program such as the one presented in this article to transform the extracted data file into a rectangular episodes table that is ready for statistical analysis.

Alternative to storing constructed variables in the EIDS, variables can be generated by the researcher at the time of creating an analysis file by using open-access extraction programs directly. This further improves transparency and replicability of studies. Extraction programs, however, are not available for all types of variables, particularly context- or database-specific variables. Additionally, such programs are not always created in a software program that is familiar to all researchers. Moreover, for very large databases constructing all variables from scratch for each research project requires a lot of computational power and time. In many cases, applying available standardized extraction programs to the EIDS can increase the use of historical demographic data stored in the IDS and the replicability of such studies.

3 STORING DATA IN THE EIDS

This section describes the EIDS and how data should be stored in such tables. As seen in Figure 1, extraction programs can be used to create the EIDS, to construct extended variables and store them in the EIDS, and to select variables from the IDS and the EIDS to create a dataset for analysis. The responsibility for creating extended variables and storing them in the EIDS lies on programmers and database administrators.

The EIDS tables have the same structure as the IDS tables. The only difference is that besides Day, Month and Year, the time stamp must also contain the column DayFrac. This field is aimed at handling date collisions which occur when there is more than one Value of a specific Type on the same date. For example, a variable indicating whether the previously born child is still alive (PrevChildIndicator) could assume three different Values: 1 (alive), 2 (dead and less than two years elapsed from the previous birth) or 3 (dead and more than two years elapsed from the previous birth). If a child is born and dies on the same day, the child's mother would have two different Values for the Type PrevChildIndicator on such date: 1 and 2. The collision that is created can be handled by adding a fraction of a day, using the column DayFrac, to the date corresponding to the Value occurring later in chronological order (in this case the Value 2). This can be thought of as a sequence number (expressed in decimals) used to define the temporary order of such Values; theoretically, this corresponds to adding some hours to a day. By using this field it is possible to sort Values in the correct chronological order when creating a dataset for analysis.

An example of an INDIVIDUAL_EXT table is shown in Table 6. It stores the individual level extended variables "Civil_status", "ChildBirth" and "PrevChildIndicator". Civil_status was constructed based on the marriage and death dates that were registered in the INDIVIDUAL table, while ChildBirth and PrevChildIndicator were constructed by selecting mother-child links from the INDIV_INDIV table and birth and death dates from the INDIVIDUAL table. The INDIVIDUAL_EXT table also contains the variables AtRisk_fertility and AtRisk_mortality, which are discussed in more detail in Section 5 of this article.

The IDS has a hierarchical structure. All individuals belong to a Context (e.g. household) and each Context can be a part of a higher level Context (e.g. town, country). This hierarchical structure is useful to construct and store variables in the EIDS. Some variables in fact relate to individuals (e.g. civil status or individual occupation), while others relate to higher levels such as the household, town, or country (e.g. household size, census information or monthly grain prices). Contextual variables can also be created from individual attributes used at a higher level (e.g. household head occupation).

All variables added to the EIDS should be constructed at the highest possible hierarchical level. For example, the variable household size can be constructed using the household context ID and can be stored in the CONTEXT_EXT table. The same can be done for all other extended variables which relate to a context. The advantage of storing contextual level extended variables in the CONTEXT_EXT table instead of assigning values of these variables directly to individuals is that the transformation of all contextual extended variables to an individual level can be made by using the same program, instead of having to write separate code for each type of variable[1]. This substantially reduces the amount of programming and computational power that is needed. The Time Stamps of the INDIV_CONTEXT and CONTEXT_EXT tables need to be used as references when assigning extended contextual level

[1] Contextual level variables must be transformed to the individual level before conducting statistical analysis. In fact from a statistical point of view, contextual variables cannot be used on the individual level unless multilevel analysis is used.

Table 6 *Example of an INDIVIDUAL_EXT table*

Id_D	Id_I	Type	Value	Day	Month	Year	Day-Frac	Source	Value_ Id_C	Date_ type
Test_DB	1148964	Civil_status	2	11	4	1825		Local_program_2		
Test_DB	1148964	Civil_status	3	15	4	1838		Local_program_2		
Test_DB	1148964	AtRisk_mortality	1	11	4	1825		Local_program_3		
Test_DB	1148964	AtRisk_mortality	0	8	11	1855		Local_program_3		
Test_DB	1237852	Civil_status	2	11	4	1825		Local_program_2		
Test_DB	1237852	AtRisk_fertility	1	11	4	1825		Local_program_5		
Test_DB	1237852	AtRisk_mortality	1	11	4	1825		Local_program_3		
Test_DB	1237852	ChildBirth		18	11	1828		Local_program_6		Event
Test_DB	1237852	PrevChildIndicator	1	18	11	1828		Local_program_4		
Test_DB	1237852	AtRisk_fertility	0	15	9	1836		Local_program_5		
Test_DB	1237852	AtRisk_mortality	0	15	9	1836		Local_program_3		
Test_DB	1237852	AtRisk_fertility	1	10	8	1837		Local_program_5		
Test_DB	1237852	AtRisk_mortality	1	10	8	1837		Local_program_3		
Test_DB	1237852	AtRisk_fertility	0	15	4	1838		Local_program_5		
Test_DB	1237852	AtRisk_mortality	0	15	4	1838		Local_program_3		
Test_DB	1378563	Civil_status	1	18	2	1853		Local_program_2		
Test_DB	1378563	AtRisk_mortality	1	18	2	1853		Local_program_3		
Test_DB	1378563	AtRisk_fertility	1	18	2	1868		Local_program_5		
Test_DB	1378563	AtRisk_fertility	0	1	12	1874		Local_program_5		
Test_DB	1378563	AtRisk_mortality	0	1	12	1874		Local_program_3		
Test_DB	1548468	AtRisk_mortality	1	7	10	1855	0.01	Local_program_3		
Test_DB	1548468	AtRisk_mortality	0	7	10	1855	0.02	Local_program_3		
Test_DB	1548468	Civil_status	1	7	10	1855		Local_program_2		
Test_DB	1567526	Civil_status	2	16	9	1851		Local_program_2		
Test_DB	1567526	Civil_status	3	3	6	1878		Local_program_2		
Test_DB	1567526	AtRisk_fertility	1	16	9	1851		Local_program_5		
Test_DB	1567526	AtRisk_mortality	1	16	9	1851		Local_program_3		
Test_DB	1567526	ChildBirth		18	2	1853		Local_program_6		Event
Test_DB	1567526	PrevChildIndicator	1	18	2	1853		Local_program_4		
Test_DB	1567526	ChildBirth		7	10	1855		Local_program_6		Event
Test_DB	1567526	AtRisk_fertility	0	26	7	1871		Local_program_5		
Test_DB	1567526	AtRisk_mortality	0	4	8	1885		Local_program_3		
Test_DB	1945568	Civil_status	1	18	11	1828		Local_program_2		
Test_DB	1945568	Civil_status	2	16	9	1851		Local_program_2		
Test_DB	1945568	AtRisk_mortality	1	18	11	1828		Local_program_3		
Test_DB	1945568	AtRisk_mortality	0	8	6	1849		Local_program_3		
Test_DB	1945568	AtRisk_mortality	1	16	9	1851		Local_program_3		
Test_DB	1945568	AtRisk_mortality	0	3	6	1878		Local_program_3		
Test_DB	1567526	PrevChildIndicator	1	7	10	1855	0.01	Local_program_4		
Test_DB	1567526	PrevChildIndicator	2	7	10	1855	0.02	Local_program_4		
Test_DB	1897563	AtRisk_mortality	1	17	11	1836		Local_program_3		
Test_DB	1897563	AtRisk_mortality	0	12	1	1852		Local_program_3		
Test_DB	1479856	AtRisk_mortality	1	1	2	1852		Local_program_3		
Test_DB	1479856	AtRisk_mortality	0	14	11	1881		Local_program_3		

variables to individuals. Table 7 shows a CONTEXT_EXT table which contains three context level extended variables: "Household_size", "Head_occupation" and "NumberOfServants". These were created based on information stored in the IDS tables.

Some contextual level extended variables cease having values prior to the date when the last individual exits from a context. In such cases, the value -1 should be assigned to the value of the extended variable on the date when it ceases having validity. For example, the Household 35891 is formed on September 16, 1851 through the marriage of a male with Id_I 1945568 and a female with Id_I 1567526. On the marriage certificate the husband is declared to be a "Farmhand" and on the date of birth of his daughter, February 18, 1853, he is declared to be a "Farmer". These two Values of the variable "Occupation" are stored in the INDIVIDUAL table (Table 1). They are also assigned to the extended household variable "Head_occupation" for the household 35891 in the CONTEXT_EXT table (Table 7). Individual 1945568 dies on June 3, 1878. On this same date the value "-1" is assigned to the extended variable "Head_occupation" for context 35891. In cases where another individual with an occupational title becomes the head of the household, such occupational title can be assigned on the date of death instead of assigning the value -1.

Table 7 *Example of a CONTEXT_EXT table*

Id_D	Id_C	Type	Value	Day	Month	Year	DayFrac	Date_type	Source
Test_DB	12345	Head_occupation	Farmer	11	4	1825		Declared	Local_program_7
Test_DB	12345	Household_size	2	11	4	1825		Declared	householdSize_program_v1
Test_DB	12345	Household_size	3	18	11	1828		Declared	householdSize_program_v1
Test_DB	12345	Household_size	2	15	9	1836		Declared	householdSize_program_v1
Test_DB	12345	Household_size	3	17	11	1836		Declared	householdSize_program_v1
Test_DB	12345	NumberOfServants	1	17	11	1836		Declared	Local_program_8
Test_DB	12345	Household_size	4	10	8	1837		Declared	householdSize_program_v1
Test_DB	12345	Household_size	3	15	4	1838		Declared	householdSize_program_v1
Test_DB	12345	Household_size	2	8	6	1849		Declared	householdSize_program_v1
Test_DB	12345	Household_size	1	12	1	1852		Declared	householdSize_program_v1
Test_DB	12345	NumberOfServants	0	12	1	1852		Declared	Local_program_8
Test_DB	12345	Head_occupation	-1	8	11	1855		Declared	Local_program_7
Test_DB	12345	Household_size	0	8	11	1855		Declared	householdSize_program_v1
Test_DB	35891	Head_occupation	Farmhand	16	9	1851		Declared	Local_program_7
Test_DB	35891	Household_size	2	16	9	1851		Declared	householdSize_program_v1
Test_DB	35891	Household_size	3	1	2	1852		Declared	householdSize_program_v1
Test_DB	35891	NumberOfServants	1	1	2	1852		Declared	Local_program_8
Test_DB	35891	Head_occupation	Farmer	18	2	1853		Declared	Local_program_7
Test_DB	35891	Household_size	4	18	2	1853		Declared	householdSize_program_v1
Test_DB	35891	Household_size	3	1	12	1874		Declared	householdSize_program_v1
Test_DB	35891	Head_occupation	-1	3	6	1878		Declared	Local_program_7
Test_DB	35891	Household_size	2	3	6	1878		Declared	householdSize_program_v1
Test_DB	35891	Household_size	1	14	11	1881		Declared	householdSize_program_v1
Test_DB	35891	NumberOfServants	0	14	11	1881		Declared	Local_program_8
Test_DB	35891	Household_size	0	4	8	1885		Declared	householdSize_program_v1

Extended variables can be described in the METADATA table. For such variables a special content of the field *Type_T* should be indicated, for example, 'INDIVIDUAL_EXT' or 'CONTEXT_EXT'. The METADATA table can be used to document the definitions and explanations of the extended variables that were constructed and included in the EIDS, thereby allowing any researcher to access such information. The METADATA table should also contain the field *Extract* in which the name of the local or open-access program used to create the specific attribute or variable can be stored. This increases transparency, reliability and replicability. An example of a METADATA table containing information that relates to Tables 1-7 is shown in Table 8.

Table 8 *Example of a METADATA table containing variables stored in IDS and EIDS*

Id_D	Type	Value	Source	Type_T	Extract	Explanation
Test_DB	Head_occupation	Definition		CONTEXT_EXT	Local_program_7	Occupation of the male household head
Test_DB	NumberOfServants	Definition		CONTEXT_EXT	Local_program_8	Number of servants living in the household
Test_DB	Civil_status	Definition		INDIVIDUAL_EXT	Local_program_2	Civil status
Test_DB	Civil_status	1		INDIVIDUAL_EXT		
Test_DB	Civil_status	2		INDIVIDUAL_EXT		
Test_DB	Civil_status	1900-01-03		INDIVIDUAL_EXT		
Test_DB	AtRisk_mortality	Definition		INDIVIDUAL_EXT	Local_program_3	Definition of the period at risk for a mortality study
Test_DB	AtRisk_mortality	0		INDIVIDUAL_EXT		
Test_DB	AtRisk_mortality	1		INDIVIDUAL_EXT		
Test_DB	AtRisk_fertility	Definition		INDIVIDUAL_EXT	Local_program_5	Definition of the period at risk for a fertility study
Test_DB	AtRisk_fertility	0		INDIVIDUAL_EXT		
Test_DB	AtRisk_fertility	1		INDIVIDUAL_EXT		
Test_DB	PrevChildIndicator	Definition		INDIVIDUAL_EXT	Local_program_4	Life status of the previously born child
Test_DB	PrevChildIndicator	1		INDIVIDUAL_EXT		
Test_DB	PrevChildIndicator	2		INDIVIDUAL_EXT		
Test_DB	PrevChildIndicator	3		INDIVIDUAL_EXT		
Test_DB	ChildBirth	Definition		INDIVIDUAL_EXT	Local_program_6	Birth of a child event
Test_DB	Birth	Definition	Birth_register	INDIVIDUAL		Birth event
Test_DB	Marriage	Definition	Marriage_register	INDIVIDUAL		Marriage event
Test_DB	Death	Definition	Death_register	INDIVIDUAL		Death event
Test_DB	Sex	Definition		INDIVIDUAL		Sex
Test_DB	Birth_date	Definition		INDIVIDUAL		Date of birth
Test_DB	Birth_location	Definition		INDIVIDUAL		Place of birth
Test_DB	Start_observation	Definition		INDIVIDUAL	Local_program_1	Dates of entry into the database
Test_DB	End_observation	Definition		INDIVIDUAL	Local_program_1	Dates of exit from the database
Test_DB	Household_size	Definition		CONTEXT_EXT	householdSize_program_v1	Number of individuals living in the household

Extraction programs to be used with the IDS can be structured modularly, through packages that generate different types of variables or sets of variables. All software that uses input from the tables of the IDS can be shared between researchers.

4 CREATING AN EPISODES FILE FOR STATISTICAL ANALYSIS

Data stored in the IDS and the EIDS follows the Entity Attribute Data Model which contains one attribute per each record. To conduct longitudinal statistical analyses, the variables that are required for the study must first be selected from the tables using extraction software. Each extraction made from the IDS and the EIDS tables can contain different variables, depending on the selection made by the researcher and on the type of study conducted. Extractions should have a consistent format. Extracted data must next be converted into a rectangular data array, also called an episodes table (see Figure 1). Episodes are spells of time during which the values of variables remain constant and at the end of which the event of interest of the study can take place. The start and end dates of the rows of an episodes table correspond to the dates when any of the variables or events included in the data extraction change value.

The construction of an episodes table can be facilitated by using input from two types of files, a *Chronicle file* and a *Variable setup file* (see Figure 1). The *Chronicle file* contains all of the variables selected for analysis, which can be a combination of individual or context level time-varying, time-invariant variables and events stored in the IDS and the EIDS tables, or created by extraction programs and added directly to this file. In the same way as the IDS tables, the *Chronicle file* should follow the Entity Attribute Data Model, containing one row per each date of change of each variable. The fields included in the *Chronicle file* are Id_I, Type, Value, Day, Month, Year, and DayFrac (see Table 9). When constructing this file all changes in the status of a variable that take place in connection to an event should also be indicated through a time-varying variable, and both of these must be included in the *Chronicle file*. For example, when making a study about marriage, a couple marrying on 1851-04-23 would have a Value 1 for the Type "Marriage" (an event) and on the same date they would have the Value "Married" on the Type "Civil_status" (time-varying variable). By including both of these variables in the *Chronicle file* it is possible to create an episodes table without any further elaborations of the data.

Table 9 *Description of the fields contained in the Chronicle file*

Name of column	Description
Id_I	Identifying number of each individual in the data
Type	Variable name
Value	Variable value
Day	Day when the variable changes value or the event takes place
Month	Month when the variable changes value or the event takes place
Year	Year when the variable changes value or the event takes place
DayFrac	Fraction of a day, which must be assigned in cases when there is more than one Value of a specific Type on the same date. This field is used to sort Values correctly in a chronological order, and therefore DayFrac should be greater for changes in Values that take place later.

The INDIVIDUAL and INDIVIDUAL_EXT tables allow storing variable Values which correspond to contexts. The attribute Value of these kinds of variables remains empty and the field Value_Id_C is used to specify the context identifier which relates to the variable. When including these types of variables in the *Chronicle file*, the Value field must be filled in with corresponding geographic information. For example, in the INDIVIDUAL table the variable "Birth_location" is stored by leaving the Value field empty, and the field Value_Id_C contains the identifier of the context in which the individual was born. When including this variable in the data extraction, the Type "Name"[2] for the context or another desired Value of a geographic nature should be selected from the CONTEXT table and added to the Value field of the *Chronicle file*. All contextual information that is necessary for the analysis should be included in the Chronicle file as different Types, by selecting and elaborating such information from the CONTEXT and CONTEXT_CONTEXT tables (e.g. municipality, region, country).

[2] In databases where location names are not unique, it is recommended to use other types of Values or to include two different variables, for example Birth_location and Birth_region, assigning for these variables the name of the municipality and the region, correspondingly.

The INDIVIDUAL and INDIVIDUAL_EXT tables also allow storing variables which are of a date nature, for example "Birth_date". For such a variable, the Value field remains empty, and the date of birth is stored in the time stamp. The Value field should also remain empty in the *Chronicle file* for date variables. The time stamp can be later assigned to all date variables when transforming the extraction into a rectangular episodes table.

The *Chronicle file* must also contain a variable defining the period in which the individual is at risk of experiencing the event of interest for the specific research study. This variable can restrict the information presented in the episodes table to only such period of validity. For example, for a study on mortality the individual is at risk of experiencing death from the date of birth or immigration until the date of death or outmigration. For a study of fertility, the period at risk is not only defined by migration, birth and death dates, but also by the age of onset and cessation of female fecundity, usually considered to be, respectively, ages 15 and 50, as well as by the date of marriage and death of the spouse if the focus is on marital fertility. The Value 1 should be indicated for these variables when the individual becomes at risk and 0 when the individual stops being at risk. Periods of gaps in the data should also be specified. These could occur, for example, if the individual leaves the study area and returns some years later. In such case the Value 0 and 1 should be indicated, respectively, on the dates of the start and end of the period of gap. An example of these variables is shown in Table 6 (INDIVIDUAL_EXT). In the example, AtRisk_mortality defines the periods during which individuals are at risk of experiencing death, while AtRisk_fertility defines the periods during which females are at risk of experiencing the birth of a child.

The purpose of the *Variable setup file* is to store information relating to each variable included in the *Chronicle file* in order to facilitate the construction of an episodes table. The *Variable Setup file* contains the fields Type, Transition and Duration. A summary description of these fields is provided in Table 10.

Table 10 *Description of the fields contained in the Variable setup file*

Name of column	**Description**
Type	Variable name
Transition	Distinguishes whether the Type is a time-varying variable that changes value at the start of the spell (Transition = Start), an event that changes value at the end of the spell (Transition = End) or a time-invariant variable (Transition = Invariant).
Duration	Distinguishes whether the Values of a Type are valid only on their date of declaration (Duration = Instant) or between a date of declaration and the next date of declaration/End_date (Duration = Continuous).

One of the main functions of the *Variable setup file* is to indicate whether a Type is a time-varying variable with Values that change at the beginning of a spell (e.g. civil status), an event occurring at the end of a spell (e.g. death) or a time-invariant variable (e.g. sex). This information allows linking variable Values correctly to dates when building the episodes table. The field Transition of the *Variable setup file* can be used to store such information, specifying the values "Start", "End" or "Invariant".

The *Chronicle file* only stores Values of variable Types at each date of declaration. For many variables the Value remains constant until the next date of change or until the last End_date for the individual. For example, the variable "Civil_status" assumes the value "Single" on the date when an individual is born, and the Value "Married" on the date of marriage. The individual is, however, single from the date of birth until the date of marriage, and is married from the date of marriage until the date of death of either of the spouses (or until the individual's last observed date). By assigning the values "Instant" or "Continuous" in the field Duration of the *Variable setup file*, it is possible to specify whether variable Values are valid only on their date of declaration or whether they are valid during the period elapsing between two different dates of declaration (or a date of declaration and the individual's End_date in the database). In the latter case, the value of a variable is assigned to all rows of the episodes table occurring after the date of change in the variable.

Extraction programs created for the IDS community should produce a *Chronicle file* and a *Variable setup file* in order to facilitate the use of the data created by such programs for research. These files in fact allow the creation of an episodes table for analysis that combines data produced using one or more extraction programs with data stored in the IDS and EIDS tables. It is much easier to combine these files than outputs from extraction programs that were already given in a rectangular format. Another advantage of using the *Chronicle file* and the *Variable setup file* is that they allow for extraction programs to be structured modularly instead of having to rewrite code to create all types of variables within the same program. For example, the program developed by Alter (Forthcoming) can be employed to create variables for a fertility analysis based on family reconstitution data. The output of such program is a *Chronicle file* and a *Variable setup file*, which can be easily combined with other variables stored in the IDS and the EIDS tables (e.g. occupation, place of birth) as well as with variables created from other extraction programs or by the researcher specifically for the study.

Table 11 shows an example of a *Chronicle file* for a mortality study based on variables from the previously shown IDS and EIDS tables. It contains the individual variables Birth_date, Birth_location and Sex, the contextual variables Household_size, Head_occupation, and NumberOfServants, the event Death and the variable AtRisk_mortality which defines the period during which the individuals are under exposure. The *Variable setup file* linked to this extraction is shown in Table 12. Table 13 shows an episodes table produced from this extraction and Table 14 instead shows an episodes table produced for a fertility study.

Table 11 *Chronicle file for a mortality study*

Id_I	Type	Value	Day	Month	Year	DayFrac
1148964	Birth_date	1804-11-8	8	11	1804	
1148964	AtRisk_mortality	1	11	4	1825	
1148964	Head_occupation	Farmer	11	4	1825	
1148964	Household_size	2	11	4	1825	
1148964	NumberOfServants	NoValue	11	4	1825	
1148964	Household_size	3	18	11	1828	
1148964	Household_size	2	15	9	1836	
1148964	Household_size	3	17	11	1836	
1148964	NumberOfServants	1	17	11	1836	
1148964	Household_size	4	10	8	1837	
1148964	Household_size	3	15	4	1838	
1148964	Household_size	2	8	6	1849	
1148964	Household_size	1	12	1	1852	
1148964	NumberOfServants	0	12	1	1852	
1148964	AtRisk_mortality	0	8	11	1855	
1148964	Death		8	11	1855	
1148964	Birth_location	Kävlinge				
1148964	Sex	Male				
1237852	Birth_date	1807-4-12	12	4	1807	
1237852	AtRisk_mortality	1	11	4	1825	
1237852	Head_occupation	Farmer	11	4	1825	
1237852	Household_size	2	11	4	1825	
1237852	NumberOfServants	NoValue	11	4	1825	
1237852	Household_size	3	18	11	1828	
1237852	AtRisk_mortality	0	15	9	1836	
1237852	AtRisk_mortality	1	10	8	1837	
1237852	Household_size	4	10	8	1837	
1237852	NumberOfServants	1	10	8	1837	
1237852	AtRisk_mortality	0	15	4	1838	
1237852	Death		15	4	1838	
1237852	Birth_location	Hög				
1237852	Sex	Female				
1378563	AtRisk_mortality	1	18	2	1853	

Table 11 continued on next page

Id_I	Type	Value	Day	Month	Year	DayFrac
1378563	Birth_date	1853-2-18	18	2	1853	
1378563	Head_occupation	Farmer	18	2	1853	
1378563	Household_size	4	18	2	1853	
1378563	NumberOfServants	1	18	2	1853	
1378563	AtRisk_mortality	0	1	12	1874	
1378563	Birth_location	Lund				
1378563	Sex	Female				
1479856	Birth_date	1831-12-15	15	12	1831	
1479856	AtRisk_mortality	1	1	2	1852	
1479856	Household_size	3	1	2	1852	
1479856	NumberOfServants	1	1	2	1852	
1479856	Head_occupation	Farmer	18	2	1853	
1479856	Household_size	4	18	2	1853	
1479856	Household_size	3	1	12	1874	
1479856	Head_occupation	-1	3	6	1878	
1479856	Household_size	2	3	6	1878	
1479856	AtRisk_mortality	0	14	11	1881	
1479856	Birth_location	Lund				
1479856	Sex	Male				
1548468	AtRisk_mortality	1	7	10	1855	0.01
1548468	AtRisk_mortality	0	7	10	1855	0.02
1548468	Birth_date	1855-10-7	7	10	1855	
1548468	Death		7	10	1855	
1548468	Household_size	4	7	10	1855	
1548468	NumberOfServants	1	7	10	1855	
1548468	Birth_location	Lund				
1548468	Sex	Female				
1567526	Birth_date	1821-7-26	26	7	1821	
1567526	AtRisk_mortality	1	16	9	1851	
1567526	Head_occupation	Farmhand	16	9	1851	
1567526	Household_size	2	16	9	1851	
1567526	NumberOfServants	NoValue	16	9	1851	
1567526	Household_size	3	1	2	1852	
1567526	NumberOfServants	1	1	2	1852	
1567526	Head_occupation	Farmer	18	2	1853	
1567526	Household_size	4	18	2	1853	
1567526	Household_size	3	1	12	1874	
1567526	Head_occupation	-1	3	6	1878	
1567526	Household_size	2	3	6	1878	
1567526	Household_size	1	14	11	1881	
1567526	NumberOfServants	0	14	11	1881	
1567526	AtRisk_mortality	0	4	8	1885	
1567526	Death		4	8	1885	
1567526	Birth_location	Kävlinge				
1567526	Sex	Female				
1897563	Birth_date	1819-8-5	5	8	1819	
1897563	AtRisk_mortality	1	17	11	1836	
1897563	Household_size	3	17	11	1836	
1897563	NumberOfServants	1	17	11	1836	

Table 11 continued on next page

Id_I	Type	Value	Day	Month	Year	DayFrac
1897563	Household_size	4	10	8	1837	
1897563	Household_size	3	15	4	1838	
1897563	Household_size	2	8	6	1849	
1897563	AtRisk_mortality	0	12	1	1852	
1897563	Birth_location	Malmö				
1897563	Sex	Male				
1945568	AtRisk_mortality	1	18	11	1828	
1945568	Birth_date	1828-11-18	18	11	1828	
1945568	Household_size	3	18	11	1828	
1945568	NumberOfServants	NoValue	18	11	1828	
1945568	Household_size	2	15	9	1836	
1945568	Household_size	3	17	11	1836	
1945568	NumberOfServants	1	17	11	1836	
1945568	Household_size	4	10	8	1837	
1945568	Household_size	3	15	4	1838	
1945568	AtRisk_mortality	0	8	6	1849	
1945568	AtRisk_mortality	1	16	9	1851	
1945568	Head_occupation	Farmhand	16	9	1851	
1945568	Household_size	2	16	9	1851	
1945568	NumberOfServants	NoValue	16	9	1851	
1945568	Household_size	3	1	2	1852	
1945568	NumberOfServants	1	1	2	1852	
1945568	Head_occupation	Farmer	18	2	1853	
1945568	Household_size	4	18	2	1853	
1945568	Household_size	3	1	12	1874	
1945568	AtRisk_mortality	0	3	6	1878	
1945568	Death		3	6	1878	
1945568	Birth_location	Malmö				
1945568	Sex	Male				

Table 12 *Variable setup file for a mortality study*

Type	Duration	Transition
AtRisk_mortality	Continuous	Start
Birth_date	Continuous	Invariant
Birth_location	Continuous	Invariant
Death	Continuous	End
Head_occupation	Instant	Start
Household_size	Continuous	Start
NumberOfServants	Continuous	Start
Sex	Continuous	Invariant

Table 13 *Episodes table for a mortality study*

Id_I	date1	date2	Head_occupation	Household_size	NumberOfServants	Birth_location	Sex	Birth_date	Death
1148964	11apr1825	18nov1828	Farmer	2	-1	Kävlinge	Male	08nov1804	0
1148964	18nov1828	15sep1836		3	-1	Kävlinge	Male	08nov1804	0
1148964	15sep1836	17nov1836		2	-1	Kävlinge	Male	08nov1804	0
1148964	17nov1836	10aug1837		3	1	Kävlinge	Male	08nov1804	0
1148964	10aug1837	15apr1838		4	1	Kävlinge	Male	08nov1804	0
1148964	15apr1838	08jun1849		3	1	Kävlinge	Male	08nov1804	0
1148964	08jun1849	12jan1852		2	1	Kävlinge	Male	08nov1804	0
1148964	12jan1852	08nov1855		1	0	Kävlinge	Male	08nov1804	1
1237852	11apr1825	18nov1828	Farmer	2	-1	Hög	Female	12apr1807	0
1237852	18nov1828	15sep1836		3	-1	Hög	Female	12apr1807	0
1237852	10aug1837	15apr1838		4	1	Hög	Female	12apr1807	1
1378563	18feb1853	01dec1874	Farmer	4	1	Lund	Female	18feb1853	0
1479856	01feb1852	18feb1853		3	1	Lund	Male	15dec1831	0
1479856	18feb1853	01dec1874	Farmer	4	1	Lund	Male	15dec1831	0
1479856	01dec1874	03jun1878		3	1	Lund	Male	15dec1831	0
1479856	03jun1878	14nov1881	-1	2	1	Lund	Male	15dec1831	0
1548468	07oct1855	07oct1855		4	1	Lund	Female	07oct1855	1
1567526	16sep1851	01feb1852	Farmhand	2	-1	Kävlinge	Female	26jul1821	0
1567526	01feb1852	18feb1853		3	1	Kävlinge	Female	26jul1821	0
1567526	18feb1853	01dec1874	Farmer	4	1	Kävlinge	Female	26jul1821	0
1567526	01dec1874	03jun1878		3	1	Kävlinge	Female	26jul1821	0
1567526	03jun1878	14nov1881	-1	2	1	Kävlinge	Female	26jul1821	0
1567526	14nov1881	04aug1885		1	0	Kävlinge	Female	26jul1821	1
1897563	17nov1836	10aug1837		3	1	Malmö	Male	05aug1819	0
1897563	10aug1837	15apr1838		4	1	Malmö	Male	05aug1819	0
1897563	15apr1838	08jun1849		3	1	Malmö	Male	05aug1819	0
1897563	08jun1849	12jan1852		2	1	Malmö	Male	05aug1819	0
1945568	18nov1828	15sep1836		3	-1	Malmö	Male	18nov1828	0
1945568	15sep1836	17nov1836		2	-1	Malmö	Male	18nov1828	0
1945568	17nov1836	10aug1837		3	1	Malmö	Male	18nov1828	0
1945568	10aug1837	15apr1838		4	1	Malmö	Male	18nov1828	0
1945568	15apr1838	08jun1849		3	1	Malmö	Male	18nov1828	0
1945568	16sep1851	01feb1852	Farmhand	2	-1	Malmö	Male	18nov1828	0
1945568	01feb1852	18feb1853		3	1	Malmö	Male	18nov1828	0
1945568	18feb1853	01dec1874	Farmer	4	1	Malmö	Male	18nov1828	0
1945568	01dec1874	03jun1878		3	1	Malmö	Male	18nov1828	1

Table 14 *Episodes table for a fertility study*

Id_I	date1	date2	Civil_status	Head_occupation	Household_size	PrevChildIndicator	Birth_date	ChildBirth
1237852	11apr1825	18nov1828	Married	Farmer	2	-1	12apr1807	1
1237852	18nov1828	15sep1836	Married		3	Alive	12apr1807	0
1237852	10aug1837	15apr1838	Married		4	Alive	12apr1807	0
1378563	18feb1868	01dec1874	Single		4	-1	18feb1853	0
1567526	16sep1851	01feb1852	Married	Farmhand	2	-1	26jul1821	0
1567526	01feb1852	18feb1853	Married		3	-1	26jul1821	1
1567526	18feb1853	07oct1855	Married	Farmer	4	Alive	26jul1821	1
1567526	07oct1855	07oct1855	Married		4	Alive	26jul1821	0
1567526	07oct1855	26jul1871	Married		4	Dead_and_less_than_two_years_elapsed_from_the_previous_birth	26jul1821	0

5 PROGRAMS FOR CONSTRUCTING FILES FOR ANALYSIS FROM THE IDS

This section presents seven different programs written for STATA which can be used to create a dataset for analysis from data stored in the IDS. These programs have been developed to be used directly by researchers, and they are very easy to run. They are developed modularly and, to a large extent, can be used independently of the others.

The open-access program **Extended IDS table maker** can be used in STATA to create the EIDS tables INDIVIDUAL_EXT, CONTEXT_EXT as well as a *Chronicle file* and a *Variable Setup file*. It creates empty tables, which can be filled in with variables constructed locally or by other extraction programs. **Household size** is an example of a program that can be used to construct extended variables at the contextual level. Using the program **Import data**, variables created by extraction programs or by locally written functions can be inserted into the INDIVIDUAL_EXT and CONTEXT_EXT tables; and information relating to such variables can be added to the METADATA table.

Variables stored in the INDIVIDUAL, INDIVIDUAL_EXT, CONTEXT or CONTEXT_EXT tables can be selected by using the program **Select Type**. The program produces an Excel file which contains the columns Type, Select and Duration and which lists each unique Type stored in the tables. The researcher can specify the value 1 in the field Select for variables that should be included in the study. Under the field Duration, the researcher should specify the value "Instant" for variables where the Value is only valid on the date of declaration, and "Continuous" for variables where the values are valid from the date of declaration until the next date of declaration or the last exit date of the individual. The Excel file should be saved after making the selection. Tables 15 and 16 show examples of the tables INDIVIDUAL_SELECT, INDIVIDUAL_EXT_SELECT, and CONTEXT_EXT_SELECT, which contain variables selected from Tables 1-7 to produce an episodes table for mortality (Table 15) and fertility studies (Table 16).

Table 15 *Tables INDIVIDUAL_SELECT, INDIVIDUAL_EXT_SELECT, and CONTEXT_EXT_SELECT for a mortality study*

INDIVIDUAL_SELECT			
Type	**Explanation**	**Select**	**Duration**
Birth	Birth event		
Birth_date	Date of birth	1	Continuous
Birth_location	Place of birth	1	Continuous
Death	Death event	1	Instant
End_observation	Dates of exit from the database		
Marriage	Marriage event		
Sex	Sex	1	Continuous
Start_observation	Dates of entry into the database		
INDIVIDUAL_EXT_SELECT			
Type	**Explanation**	**Select**	**Duration**
AtRisk_fertility	Definition of the period at risk for a fertility study		
AtRisk_mortality	Definition of the period at risk for a mortality study	1	Continuous
ChildBirth	Birth of a child event		
Civil_status	Civil status		
PrevChildIndicator	Life status of the previously born child		
CONTEXT_EXT SELECT			
Type	**Explanation**	**Select**	**Duration**
Head_occupation	Occupation of the male household head	1	Instant
Household_size	Number of individuals living in the household	1	Continuous
NumberOfServants	Number of servants living in the household	1	Continuous

Table 16 *Tables INDIVIDUAL_SELECT, INDIVIDUAL_EXT_SELECT, and CONTEXT_EXT_SELECT for a fertility study*

INDIVIDUAL_SELECT			
Type	**Explanation**	**Select**	**Duration**
Birth	Birth event		
Birth_date	Date of birth	1	Continuous
Birth_location	Place of birth		
Death	Death event		
End_observation	Dates of exit from the database		
Marriage	Marriage event		
Sex	Sex		
Start_observation	Dates of entry into the database		
INDIVIDUAL_EXT_SELECT			
Type	**Explanation**	**Select**	**Duration**
AtRisk_fertility	Definition of the period at risk for a fertility study	1	Continuous
AtRisk_mortality	Definition of the period at risk for a mortality study		
ChildBirth	Birth of a child event	1	Instant
Civil_status	Civil status	1	Continuous
PrevChildIndicator	Life status of the previously born child	1	Continuous
CONTEXT_EXT SELECT			
Type	**Explanation**	**Select**	**Duration**
Head_occupation	Occupation of the male household head	1	Instant
Household_size	Number of individuals living in the household	1	Continuous
NumberOfServants	Number of servants living in the household		

Selected individual variables can be automatically added to the *Chronicle file* using the program **Append individual variables**. This program obtains the variables selected by the user from the INDIVIDUAL or the INDIVIDUAL_EXT tables and appends such information to the *Chronicle file*. It also appends information relating to these variables to the *Variable setup file*. When appending variables which correspond to contexts and which were stored in the INDIVIDUAL or INDIVIDUAL_EXT table by using the field Value_Id_C and leaving the Value field empty (e.g. "Birth_location"), the program assigns the names of the context to the attribute Value of the *Chronicle file*, obtaining this information by selecting the Type "Name" from the CONTEXT table[3].

The program **Append contextual variables** can be used to transform contextual extended variables selected for analysis into individual extended variables and to append these transformed variables to the *Chronicle file*. Information relating to these variables is also appended to the *Variable setup file.* Contextual level variables stored in the CONTEXT table can be added to the extraction and setup files using the same program. All changes in the Values of selected contextual variables which occur on or after the individual entered the context are assigned to the individual on such date of change. In many cases there is no change in the Value of a Type that occurs on the same date when the individual enters the context (Start_date). In such cases, the Value of the Type which was declared on a date that preceded the Start_date of the individual in the context is assigned on the Start_date. Only Values of Types for which the option Duration was set to "Continuous" are assigned[4]. If an individual leaves a context for some time and later returns, the Value of the latest change in a variable Type preceding the date of return into the context is assigned to the individual on such date[5].

The largest and most important program presented in this work is the **Episodes file creator**, which produces rectangular episodes files. Using input from the *Chronicle file* and *Variable Setup file*, this program combines the variables included in such extraction, transforming the extraction into a rectangular table and formatting this file to be ready for statistical analysis. The program is generic and can be used for any data extraction. Although the program has to be used in STATA, the output produced by it can be easily exported in order to conduct statistical analysis using other software.

When rectangularizing variables which correspond to dates and which were included in the *Chronicle file* by leaving the attribute Value empty, the program assigns the time stamp to the corresponding cells of columns containing such date variables. Invalid or incomplete dates are left blank in the episodes table.

In the IDS and the EIDS tables, all variable Values are stored as text, which means that variables are also stored as text in the *Chronicle file*. Even if all information is stored as text, some variables are numerical. When producing the final rectangular file, the **Episodes file creator** reformats such variables to numbers. After reformatting variables, the **Episodes file creator** copies down all Values of each Type, with the exception of Types that were distinguished as being events (Transition = End in the *Variable setup file*) or as being only valid on their date of declaration (Duration = Instant in the *Variable setup file*). At the end of this step the program replaces any remaining missing values with -1 (with the same exceptions).

As explained earlier, the *Chronicle file* must contain a variable defining the period in which the individual is at risk of experiencing the event of interest for the specific research study. The **Episodes file creator** produces an episodes table which only contains information for such period of validity. All other rows are deleted from the table.

The **Episodes file creator** also allows the user to assign labels to the Values of Types that are categorical

[3] In its current version the program does not allow to add other geographic information obtained through a linkage to the CONTEXT_CONTEXT table. Such data should therefore be included manually or using other programs.

[4] For example, the individual 1378563 enters context 35891 on February 18th, 1853. On such date there is no declaration of the variable NumberOfServants. The latest declaration of a Value of such Variable (1) was February 1st, 1852. The Value 1 is therefore assigned to individual 1378563 on February 18th, 1853.
The individual 1945568 enters context 12345 on November 18th, 1828, and there is no declaration of the variable Head_occupation on such date. The previously declared value of such variable occurred on April 11th, 1825. Since the field Duration is set to Instant for this variable, no Value is assigned to individual 1945568 on November 18th, 1828.

[5] Individual 1237852 is present in context 12345 from April 11th, 1825 until September 15th, 1836 and also from August 10th, 1837 until April 15th, 1838. From September 15th, 1836 until August 10th, 1837 she is absent from the studied area. The first declaration of the Type NumberOfServants for this context takes place on November 17th, 1836, when there is one servant. The Value 1 is assigned for such variable to individual 1237852 on August 10th, 1837, when she returns to context 12345.

and numerical. A separate file containing labels stored using the fields Type, Value and ValueLabel can be given during input. Table 17 shows an example of labels to be added to the Values of the Types Civil_status and PrevChildIndicator.

Table 17 *External table containing labels for variable values*

Type	Value	ValueLabel
Civil_status	1	Single
Civil_status	2	Married
Civil_status	3	Widow/er
PrevChildIndicator	1	Alive
PrevChildIndicator	2	Dead and less than two years elapsed from the previous birth
PrevChildIndicator	3	Dead and more than two years elapsed from the previous birth

6 CONCLUSIONS

The IDS was developed to provide a common structure for storing and sharing longitudinal demographic data obtained from historical sources, primarily family reconstitutions and population registers. One of the main aims behind the introduction of such a structure is to reduce the complexity of the use of this type of data, therefore expanding the range of potential users. The IDS and the EHPS-net also have the aim of developing common software. Although how to enter data into the IDS has been explained in detail previously, not much has been written about how to extract data from the IDS to construct files for analysis.

This article has proposed an extension to the IDS, the EIDS, which can store not only information obtained directly from the sources but also variables constructed from such data. It has also proposed a format for data extractions and for the output produced by extraction programs, consisting of a *Chronicle file* containing all the declarations of variables and events and their values, and a *Variable setup file* describing the information contained in such extraction. In addition, this work has described how extended variables could be created and data in the EIDS should be stored, how to select data from the IDS and the EIDS tables and how to create a rectangular episodes file that is ready for analysis, presenting seven programs written for STATA to conduct such steps. The solutions and programs presented can be used to extract a dataset for analysis from databases created from population registers or family reconstitutions and linked to any type of research question dealing with longitudinal analysis, as long as the data has already been transformed into the IDS.

The use of the EIDS further increases the transparency and replicability of studies that employ longitudinal historical demographic sources, as well as the coherence between research conducted by different scholars using the same databases. The format for output proposed in this work and the open-access programs presented allow the construction of datasets for analysis directly from the IDS, further expanding the range of potential users of these databases as well as the scope of historical demographic research in general.

ACKNOWLEDGEMENTS

I would like to thank George Alter, Clas Andersson and Kees Mandemakers for their advice and help and Anders Brändström, Hilde Leikny Sommerseth, Annika Westberg and two anonymous reviewers for their comments. I am also grateful for the travel funding received from the EHPS-net and for administrative assistance from Marja Koster and Paul Puschmann.

REFERENCES

Alter, G. (Forthcoming). *Birth_intervals ver 5: A database program to create episodes for birth interval analysis from vital events.* (5th ed.)

Alter, G. & Mandemakers, K. (2014). The Intermediate Data Structure (IDS) for longitudinal historical microdata, version 4. *Historical Life Course Studies,* 1, 1-26.

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.

Bengtsson, T., Campbell, C. & Lee, J. (2004). *Life under pressure: Mortality and living standards in Europe and Asia, 1700-1900*. Cambridge, Massachusetts: MIT Press.

Bengtsson, T. & Dribe, M. (2014). The historical fertility transition at the micro level: Southern Sweden 1815-1939. *Demographic Research,* 30(17), 493-533.
DOI: 10.4054/DemRes.2014.30.17

Bengtsson, T., Dribe, M., Quaranta, L. & Svensson, P. (2014). *The Scanian Economic Demographic Database. version 4.0 (machine-readable database)*. Lund: Lund University, Centre for Economic Demography.

Bengtsson, T. & Lindström, M. (2003). Airborne infectious diseases during infancy and mortality in later life in southern Sweden, 1766-1894. *International Journal of Epidemiology,* 32(2), 286-294.
DOI:10.1093/ije/dyg061

Bengtsson, T. & van Poppel, F. (2011). Socioeconomic inequalities in death from past to present: An introduction. *Explorations in Economic History,* 48(3), 343-356.
DOI:10.1016/j.eeh.2011.05.004

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B (Methodological)*, 34(2), 187-220.

Danell, C. (1981). The demographic data base of Umeå University. In: A. Brändström, & J. Sundin (Eds.), *Tradition and transition: Studies in microdemography and social change* (pp. 241-254). Umeå: Demographic Database.

Dribe, M., Van Bavel, J. & Campbell, C. (2012). Social mobility and demographic behaviour: Long term perspectives. *Demographic Research,* S10(8), 173-190.
DOI: 10.1111/1468-0297.00288

Edvinsson, S. (2000). Sweden - Umeå - the Demographic Data Base at Umeå University: A resource for historical studies. In: P. Kelly Hall, R. McCaa & G. Thorvaldsen (Eds.), *Handbook of international historical microdata for population research* (pp. 231-248). Minneapolis, Minnesota: Minnesota Population Center.

Lindeboom, M., Portrait, F. & van den Berg, G. J. (2010). Long-run effects on longevity of a nutritional shock early in life: The Dutch potato famine of 1846–1847. *Journal of Health Economics,* 29(5), 617-629.
DOI: 10.1016/j.jhealeco.2010.06.001

Lundh, C. & Kurosu, S. (2014). *Similarity in difference: Marriage in Europe and Asia, 1700-1900*. Cambridge, Massachusetts: MIT Press.

Mandemakers, K. (2000). Netherlands - Historical Sample of The Netherlands. In: P. Kelly Hall, R. McCaa & G. Thorvaldsen (Eds.), *Handbook of international historical microdata for population research* (pp. 149-178). Minneapolis, Minnesota: Minnesota Population Center.

Öberg, S. (2014). *Social bodies: Family and community level influences on height and weight, southern Sweden 1818-1968*. Bohus: Ale Tryckteam.

Quaranta, L. (2011). Agency of change: Fertility and seasonal migration in a nineteenth century alpine community. *European Journal of Population,* 27(4), 457-485.
DOI: 10.1007/s10680-011-9241-2

Quaranta, L. (2013). *Scarred for life: How conditions in early life affect socioeconomic status, reproduction and mortality in southern Sweden, 1813-1968*. Lund: Media-Tryck, Lund University.

Quaranta, L. (2014). Early life effects across the life course: The impact of individually defined exogenous measures of disease exposure on mortality by sex in 19th- and 20th-century southern Sweden. *Social Science & Medicine,* 119, 266-273.
DOI: 10.1016/j.socscimed.2014.04.007

Quaranta, L. (Forthcoming). STATA programs for using the Intermediate Data Structure (IDS) to construct files for statistical analysis. *Historical Life Course Studies.*

Reher, D. S. & Sanz-Gimeno, A. (2007). Rethinking historical reproductive change: Insights from longitudinal data for a Spanish town. *Population and Development Review,* 33(4), 703-727. DOI: 10.1111/j.1728-4457.2007.00194.x

Schofield, R. G., Reher, D. & Bideau, A. (Eds.). (1991). *The decline of mortality in Europe*. Oxford: Clarendon Press.

Stead, W., Hammond, W. & Straube, M. (1982). A chartless record - is it adequate? In: *Proceedings of the annual symposium on computer application in medical care* (p.89-94). *American Medical Informatics Association.*
DOI: 10.1007/BF00995117

Therneau, T. M. & Grambsch, P. M. (2000). *Modeling survival data: Extending the Cox model*. New York: Springer-Verlag.

Tsuya, N. O., Feng, W., Alter, G. & Lee, J. Z. (Eds.). (2010). *Prudence and pressure: Reproduction and human agency in Europe and Asia, 1700-1900*. Cambridge, Massachusetts: MIT Press.

Van Bavel, J. (2004). Deliberate birth spacing before the fertility transition in Europe: Evidence from nineteenth-century Belgium. *Population Studies,* 58(1), 95-107.
DOI: 10.1080/0032472032000167706