Defining and Distributing Longitudinal Historical Data in a General Way Through an Intermediate Structure

George Alter, Kees Mandemakers, Myron Gutmann

Abstract: Articles longer than six pages have an English abstract outlining shortly the article's content and composition. The abstract's length should not exceed 150 words. Please enclose the German version of the abstracts. Both versions are going to be published online in our abstract-database (HSR-RETROSPECTIVE, www.hsr-retro.de).

German Title:

Zusammenfassung:

Keywords: Data Management • International Comparative Research • Accessing Data • Sampling

1. Introduction

In recent years, the study of population history has shifted from studying 'demographic regimes' and large-scale processes to analysing longitudinal micro data in the form of 'life histories'. Because demography deals with the fates and choices of individuals, the micro-level is optimally suited to study chains of causation. Thus, demographers can now analyse processes of family formation and change that span life times and can even be followed across generations. New strategies of data collecting and data sharing, as well as new statistical techniques, have opened up vistas of research that currently reshape historical demography's landscape (Settersten 2002). More broadly, the life course approach expands the whole field of social history providing 'a framework for studying phenomena at the nexus of social pathways, developmental trajectories, and social change' (Elder et al. 2003, 10).

The various databases constructed around micro-data have stimulated and rejuvenated the field of family history. So far, work on databases with longitudinal information on individuals and their families has been localized, only rarely covering an entire country (Kelly Hall et al. 2000). The logical next step in this scientific development is *comparing life courses across local and national databases*. One of the pioneering projects in this field is the Eurasia Project, which compared the life courses of historical populations in Belgium, Sweden, Italy, Japan and China. In *Life Under Pressure* (Bengtsson et al. 2004), the group studied mortality across family systems, revealing differences in internal redistribution of food, differential protection in times of economic stress, and different power relations between generations and sexes. When more datasets in different parts of the world can be made comparable, it will become feasible to study differences in family life, in family ties and in individual behaviour by religion, by level of urbanization and economic specialization, by system of communal support et cetera. Understanding variation and different responses to similar economic conditions or processes (modernisation, globalisation) will provide an important historical perspective on present-day challenges.

Everyone has been impressed by the success of the I*ntegrated Public Use Microdata Series Project* (IPUMS; Ruggles et al. 2008) in encouraging new research with historical data. By providing data in a consistent and easy to use form, IPUMS has generated thousands of studies with data that were already available in less user-friendly versions. Those of us who collect and work with longitudinal historical data cannot help but wonder whether something similar can be done with the sources that we use. We now have an embarrassment of potential riches, including classic data sets, such as Henry's data for France and John Knodel's German villages, and on-going projects on both sides of the Atlantic, Japan, China, and elsewhere. These data can shed light on a wide range of questions about demographic and family patterns, social mobility, and

other issues, but they are used by a very small community of researchers. The obvious question is: Can longitudinal historical databases follow the lead of IPUMS and make data sets available to a much wider community of researchers?

In this article, we discuss some of the challenges of longitudinal historical data: selection, fuzziness, censoring (Section 2). Section 3 focuses on practical problems by contrasting historical longitudinal data with contemporary longitudinal surveys and explaining the need for rectangular datasets. Section 4 proposes a strategy to simplify the sharing of historical longitudinal data: an intermediate data structure (IDS). We end with some thoughts on the benefits of the IDS approach and activities that will complement this approach.

This article is a result of several workshops discussing these problems. The first one was a meeting at Montréal in November 2003, in which the issue of the problematic character of longitudinal historical data was raised (Dillon and Roberts 2006). The second workshop, titled 'Disseminating and Analyzing Longitudinal Historical Data', took place at Amsterdam in February 2006.[1] Although the participants in the Amsterdam meeting recognized the complex nature of their longitudinal databases, the workshop ended with a consensus on how to make progress. First, it was agreed that standardization in the products of the different databases should help the researcher enormously. Second, an intermediate structure was proposed that could mediate between the original databases and the data sets required for analysis. On May 1-2, 2008, the Inter-university Consortium for Political and Social Research (ICPSR) hosted a planning group to continue working. This resulted in a model for data sharing, which was presented to an open meeting of historical databases at the Social Science History Association meeting in Miami, October 22, 2008.[2] This article describes the proposal that emerged from that planning meeting.

---

[1] Part of the workshop was the publication of Questionnaires with key information about the databases the participants were representing, including Historical Database of the Liège Region (Belgium), Scania Database (Sweden), Registre de la population du Québec ancien (PRDH), Historical Sample of Flanders, Demographic Database Umeå (Sweden), Victorian Scotland database, Connecticut River Valley Project (USA), Texas Longitudinal Data Project (USA), Migration Database (USA, based on genealogies), Danjuro Database Japan, Historical Sample of the Netherlands (HSN), Koori Health Research Database (KHRD) 1855-1930 (Australia), Melbourne Lying-In Hospital Cohort: 1857-1900, Utah Population Database, Geneva Database, IPUMS database (census USA), Norwegian Census Database, see http://historicaldemography.net/questionnaires.php

[2] This work was made possible by grants from the Netherlands Organization for Scientific Research (NOW), Humanities (Internationalisation, 236-053-004)), the Inter-university Consortium for Political and Social Research (ICPSR) and the Demographic Data Base at Umea University in Sweden (DDB).

2. Challenges of Historical Longitudinal Data

We see a number of interconnected problems that prevent researchers from using longitudinal historical data. Some of these problems are due to the dynamics of populations that change over time. Recent research has also emphasized the multi-level and relational aspects of longitudinal data. Lives are lived within concentric circles of families, households, kin networks, communities, etc. Time and context create conceptual and practical problems, some of which are unique to historical sources.

2.1 Thinking About Time

Most social scientists, whether they come from history, sociology, economics, or other disciplines, are not trained to conceptualize processes that develop over time. Training in demography is the most conducive to thinking in a longitudinal perspective, but even in demography the life table is usually presented as a way of mediating between a *life course perspective* (expectation of life) and essentially *cross-sectional data* (census counts and vital registration). Longitudinal data offer powerful opportunities for designing tests of hypotheses, but they also have serious pitfalls for the unwary. Under the heading of opportunities, we would emphasize the *importance of viewing life histories sequentially and asking how prior events affect decision making*. For example, a classic issue in fertility research is the replacement effect. Did couples who would otherwise have practiced family limitation resume having children following the death of a child? The sequence of events is critical here. Since short birth intervals can contribute to infant mortality, we must look at fertility following a child death, rather than comparing fertility rates of couples with and without child deaths. Moreover, we must distinguish between infant deaths and the deaths of children above age one, because the termination of breastfeeding after an infant death increases fertility (see Alter 1988).

Every life history incorporates multiple time dimensions. Demographers often refer to the *trinity of age, period, and cohort*. The effects of historical events (e.g. wars, famines, epidemics) may differ by age or by cohort. Moreover, there are many versions of each of these time scales. Age may be time since birth, but it may also be time since some other event, such as marriage or leaving school. Family reconstitution studies are often organized by marriage cohorts, which can combine several decades of birth cohorts (compare Kok et al. 2005). The experiences of siblings may differ in important ways, because they are born at different stages of the family life cycle and experience critical events at different ages.

2.2 Selection

Among the pitfalls inherent in longitudinal data are the many forms of selection. *Selection occurs any time individuals differ in their propensities to experience an event or transition*. When a population is followed over time its composition changes as those with higher propensities experience transitions (death, childbirth, migration, etc.) and move to different statuses. If a population starts out with equal numbers of "*movers*" and "*stayers*," it is bound to end up with a higher proportion of "stayers" than "movers." Changes in composition due to selection are easily confused with intentional behavior. For example, average birth intervals tend to get shorter as birth order increases. One might be tempted to infer that this pattern is due to differences in family size preferences, but it will occur without any differences in behavior. It requires only differences among women in fecundity. All women are at risk of first births, but only women who have short birth intervals will be able to have eight, nine, or ten births before reaching menopause. Since women who tend to have longer birth intervals are unlikely to reach higher parities, the proportion of women with long intervals decreases and average birth intervals become shorter as parity increases.

Some apparently simple computations become complicated with longitudinal data. Consider the problem of computing average ages at first and last birth. When life histories are incomplete ("censored") the subpopulation available for computing age at first birth will be different from the subpopulation available for computing age at last birth. The former requires that women are under observation from the date of their marriage. The latter requires that women are under observation until they reach approximately age 50. Married women who migrate into the study area will fail on the first test, and women who move out of observation fail the second test. In highly mobile populations, like growing cities, the geographically stable subpopulation is not representative of the population as a whole.

2.3 Informative Censoring

Selection is unavoidable in longitudinal data, and it points to a second major problem, informative censoring. *Informative censoring occurs when the probability that a life history will end is correlated with the probability of the event that we wish to measure*. So, if we are studying child mortality, the incomplete life histories of children who are more likely to die should not be systematically shorter or longer than those of children who are less likely to die. Informative censoring is a serious problem in databases constructed from "passive registration systems." In these systems we only know that an individual is present and under observation when an event occurs. *Individuals who leave observation do not announce their departures, so we do not know how long*

they were at risk between the last recorded event and their unobserved departures. This is the classic problem posed by parish registers. The parish registers tell us when births, marriages, and deaths occurred but not when people moved out of the parish. In the absence of censuses, we cannot construct birth or death rates, because we do not know how many people were at risk of dying. Louis Henry solved this problem by developing the rules of family reconstitution, which strictly limit the types of events that can be used to close observation in a life history. Life histories ending with an event providing information about the event of interest are excluded from the analysis. For example, the death of a sibling cannot be used to close observation of life histories used to analyze child mortality. Since siblings share the same home environment, the mortality of siblings tends to be correlated, and children whose siblings died are more likely to die themselves. If we do not know whether a specific child lived or died, we cannot use the death of a sibling to determine the length of time that subject survived. When the life histories of children with unobserved deaths are censored by the deaths sibling, the level of child mortality is overestimated.

2.4 Fuzzy Dating

Longitudinal data analysis implies that most of the variables are exactly dated. This is not always the case. We often work with sources in which transitions are not recorded. For example, we may know that an individual was a lawyer in 1850 and a judge in 1860 but not know when the change occurred. *It helps to identify three types of dating:*

1) *Dated Events.* These are life-course transitions for which exact dates are available. Typically, we have exact dates for demographic data like birth, death and marriage from civil registration or parish registers.
2) *Dated Declarations.* We often know that a person was in a certain state at a specific time without knowing when that status began. For example, censuses usually report marital status and occupation, but they do not give the date of marriage or how long the occupation has been practiced. (Alter and Gutmann 1999, 171).
3) *Interval Censored Transitions.* It is often the case that we know a transition occurred within a certain period of time. For example, we may know that a person migrated between 1880 and 1890 without knowing the exact date of migration.

The latter two types of dating require strategies for interpolating information within life histories. How long can we attribute an occupation before and after it has been reported? If we observe different occupations in successive censuses, should we try to assign a date to the transition?

In some cases the period of uncertainty can be shortened by using information from related individuals. For example, we may observe that a migrant was recorded in a population register on a line between two births. This reduces the range of uncertainty, but it makes inferring dates a complicated and tricky business.

It is also common to find incomplete dates. The year may be given, but not the month or day. Ages locate dates of birth within a year. Different kinds of analysis require different levels of precision. Monthly data may be required for analyzing infant mortality or fertility. Occupational mobility may be analyzed in broad age groups.

2.5 Multi-Level and Relational Data

In a classic study Tamara Hareven (1982) evoked the differences between family time and industrial time. Longitudinal data allow us to examine the intersections between individual life histories and many other time dimensions. Many of these databases span long periods of time, which allow us to reconstruct kinship networks extending backward in time and outward to many degrees of kinship. Information about the lives of near and remote kin create opportunities for research on questions on the boundaries between genetic and social science research. Population registers provide household as well as kinship information. We can also include information on conditions and events at the neighbourhood, community, and national levels.

Our research questions often involve events in the life histories of related individuals. For example, customs often dictated that daughters should marry in the birth order, and Daniel Scott Smith has argued that strict adherence to this rule is a sign of strong parental power over children (Smith 1973). To create a variable like "number of unmarried elder sisters" we must identify each woman's older sisters and keep track of when they married. Linking histories among individuals and across levels can be very rewarding, but it can also be technically challenging.

We use "multi-level" to refer to the many contexts in which individuals interact and share experiences. A basic list of levels may include:

1) *Individual:* genetic attributes, life history describing date of birth and death,
2) *Family:* characteristics of parents, siblings, spouse, children
3) *Household:* description of the residential group including kin and non-kin
4) *Community:* local institutions (e.g. welfare and social support), environment, industrial structure, population density
5) *Region:* economic opportunities, prices of commodities
6) *Nation:* legislation and policies on taxation, subsidies, welfare, etc.
7) *International:* wars, epidemics

Data at all levels are time-varying, which means that events at each level must be coordinated with the timing of each individual biography.

We consider our databases "relational," because they provide links between pairs of individuals that can be used to reconstruct broader networks. For example, all kinship networks can be reconstructed from two basic relationships: parent-child and husband-wife. A sibling is a "parent's child," and a first cousin is a "parent's parent's child's child." Kinship relations can be made more precise by specifying gender ("mother's mother's son's son") and birth order ("mother's first son").

Kin networks can be conceptualized as time-varying attributes of individuals. Instead of taking a static genealogical approach to kinship, we can develop measures that capture the number and types of kin available at any moment in time. This would mean that a subject's kin network would expand when she marries, for example, and children would be counted differently than adults.

Households can also be conceptualized as relations. A household is a collection of individuals who share a residence at a moment in time. Thus, each life history can be linked to a sequence of residences that are delimited in time. This perspective lends itself to descriptions of household composition focused on each individual rather than measures of household structure, which are usually constructed from the perspective of the head of household. For example, it is much different to ask whether a subject's father was present in the household than to ask whether a subject is the child of the household head.

3 Distributing Historical Longitudinal Databases

3.1 Comparing Historical and Contemporary Longitudinal Databases

It is useful to consider the contrasts between historical and contemporary longitudinal databases on these issues. The longitudinal data sets that play an important role in contemporary demographic, economic, and social research differ in three important ways from the historical data bases:

1) *Since most of them are based on surveys, informative censoring is not a fundamental problem*. Individuals are censored at the date of the most recent survey. These surveys do have a problem with individuals who are lost from one wave to the next. Sample attrition and non-responses are essentially forms of informative censoring, and we may be able to learn from strategies used to handle these problems.
2) *Most longitudinal surveys are based on panel designs rather than continuous observation*. Many researchers analyze these data as panels, i.e. linked

cross-sections. Time is simplified by collapsing all durations into the intervals between panels. Many of the details available in continuous time are lost, but there is usually not much difference between analyses done in discrete and continuous time. Panel data is inherently rectangular (one row per subject per panel), so it can be processed by statistical software directly.

3) *Contemporary longitudinal data sets tend to be broad but shallow, while historical data sets tend to be narrow but deep*. We mean by this that contemporary data sets have a large number of variables because it is relative easy to collect them, but they cover relatively short periods of time. In contrast, historical data sets tend to have a small number of variables but often cover long periods of time. Researchers working with contemporary surveys can often choose among hundreds of questions asked in each panel. Contemporary surveys contain questions about health, income, wealth, attitudes, and many other subjects. Most of these types of information are unavailable historically, although some can be created from multilevel and relational data or by adding sources like tax registers. In comparison, the sparsely described biographies derived from historical sources must be massaged to create the relevant variables for our analyses.

3.2 Rectangularization

Longitudinal data must be converted into a rectangular data array before it can be analyzed by standard statistical packages. This is a purely technical problem, but one with important implications. Life histories are anything but rectangular. Individuals can marry several times or not at all. They can have zero to twenty offspring. They can migrate or change addresses and occupations many times. Each of these contingencies must be translated into rows and columns in a data matrix for statistical analysis.

This process is further complicated when we want to use *time-varying covariates*. The standard way of constructing time-varying covariates is to divide each life history into a sequence of time intervals, so that every covariate takes only one value during each interval of time. Some statistical packages have facilities for splitting intervals. For example, a life history can be split into two rows at the date of marriage to separate time spent unmarried from time spent married. Creating time-varying covariates in a statistical package can become very tricky, however:

1) It often involves moving between time dimensions, such as age and calendar time.
2) It assumes that each record contains all the information about all possible changes in time-varying covariates.

The most difficult challenge is capturing changes over time in covariates describing other individuals, such as household composition. If household composition matters, we need to start a new interval every time a person enters or leaves the household. If we are interested in age composition, we also need to start a new interval whenever any person in the household crosses a boundary between age groups. The events that change covariates may not even occur within the same household. For example, we may want to know how many of the subject's siblings are married, even if they are living elsewhere.

The underlying structure of a historical database is often very different from the format of the dataset needed for analysis. Since historical databases are often constructed by linking several types of records (vital events, censuses, tax registers, etc.) together, the database may be structured to reflect these documents, or it may combine the data into individual life histories or some other format. In either case, the database will be relational, and the links between records are essential information. In contrast, the datasets used for analysis are usually rectangular, and relational information (households, kinship networks, etc.) must be represented by summary measures. There is presently no standard way of preparing data for longitudinal analysis, and every research project using historical data has developed its own unique computer programming.

Figure 1: Two Uneconomic Strategies Collecting Data for Scientific Research from Historical Longitudinal Databases

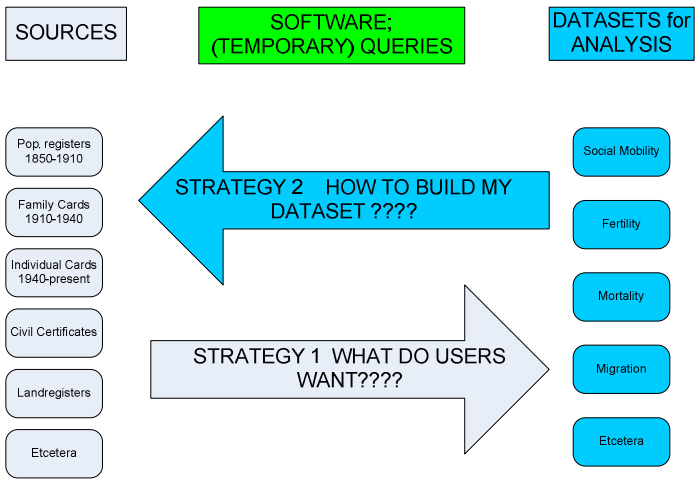


Figure 1 shows the two ways that data is usually extracted from a database for analytical purposes:

1) The database administrator opens all or parts of the database to the user, and the user builds a dataset structured to answer his/her specific research questions.
2) The researcher explains his or her data needs to the database administrator, who creates the required datasets, sometimes using previously created programs or datasets.

Both ways have several disadvantages. The most important are the most obvious:

1) Every research question requires its own dataset which means that a lot of effort must be put into each dataset.
2) Both approaches risk misinterpretation, because the researcher may misunderstand an essential aspect of the data or the database administrator may misunderstand the research question.
3) The second approach also places a financial burden on the database in question. Time used for the creation targeted datasets is taken away from time spent on developing the database itself.

All of this implies that the users of historical longitudinal data require an array of conceptual and technical skills that must appear daunting to all but the most dedicated and/or foolhardy graduate students. Less obvious but very important is the way that both strategies restrict the number of researchers who can access the data. It is time-consuming to go to Umeå or Salt Lake City, and cumbersome to exchange complex communications with database administrators from a distance. Only experienced researchers with funding will take this step and even they will hesitate to make comparative analyses based on several databases with longitudinal data. How do we encourage new researchers to enter the field? We do not have answers, but we think the use of an Intermediate Data Structure (IDS) is a strategy that may contribute to a solution.

4 Intermediate Data Structure (IDS)

4.1 Overview of the IDS

Figure 2: Strategy with Intermediate Structure Collecting Data for Scientific Research from Historical Longitudinal Databases.

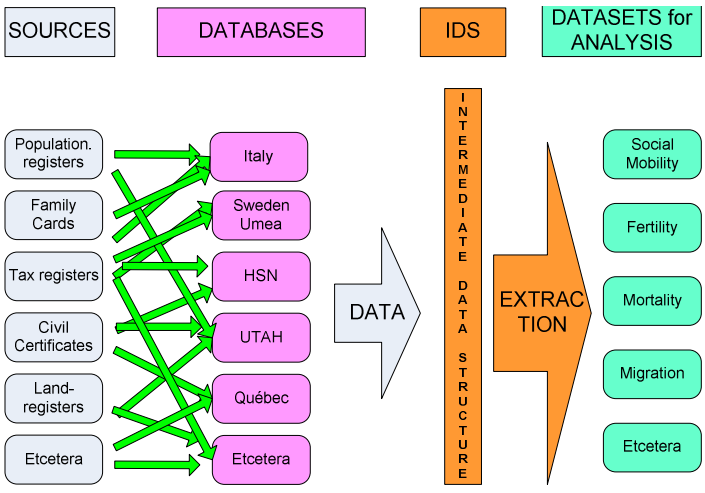


Figure 2 presents a new strategy based on an Intermediate Data Structure (IDS). The basic idea is that all relevant longitudinal databases will transfer their data into a simple common data format. The format of this data structure must be specified by the community of users. On the left side of the diagram are the various types of sources included in historical longitudinal databases. These sources vary widely from baptisms, marriages, and burials in parish registers to medical examinations and payment histories in pension records. Each database captures and stores data in a different way, and it is impossible to create a single data management structure that will work for every situation. On the right side of the diagram are the data files that researchers require for analysis. These files should be in a rectangular format that will be compatible with standard statistical packages (SPSS, SAS, Stata, etc.). While some statistical packages can manage hierarchical or relational file structures, these complexities impose costs on the user and limit accessibility. Between the sources and the analytical formats is an Intermediate Data Structure (IDS), which provides a standard format for all databases. The IDS requires two kinds of computer programs:

1) *Data transfer*. Data must be reformatted for transfer from the database to the IDS. This includes original data as well as enhancements and standardizations, such as recoding occupations into the HISCO system. Transferring information from the source database into the IDS format also implies the generation of descriptive metadata to document the source and construction of all data. Since each source database is unique, this process will vary in many details. This approach gives each database control over what and how they disseminate their data.
2) *Extraction.* The extraction process moves data from the IDS into file formats designed for analysis. Since the requirements of every type of analysis differ (fertility, mortality, social mobility, etc.), we expect to have many specialized extraction programs. However, all extraction programs will start with the IDS, and they will work on any dataset that includes the necessary attribute types. Extraction programs will be modular, and some types of analysis will require "workflows" that link together several extraction services. This process creates standardized information for all databases.

This approach separates the programs that transfer data from the original database into the IDS from the programs that create datasets in the rectangular format used by statistical packages. All databases will have the same structure, which will be independent of the form in which they were originally captured or stored. Researchers will not need to learn a new set of formats and relational structures for every database. Consequently, data extraction programs can be re-used and adapted to other purposes, and the steps involved in preparing data

for analysis will be more open and transparent. Each database providing data will be responsible for transferring their data into the IDS, and databases will be able to choose how their data are represented in the IDS to control how it can be used.

4.2 Principles

1) The database consists of two kinds of entities, persons and contexts, and the relations among persons and between persons and contexts.
2) Identifying unique persons from multiple appearances in the sources (record linkage) must be done by the data producer.
3) Contexts locate individuals in physical and social space. Contexts are multidimensional and may be nested.
4) The links between individuals and contexts tell us who lived together and who shared the same environments and experiences.
5) All entities in the IDS can be located in time. A Time Stamp is used to date to all attributes of persons and contexts. Time stamps must be constructed by the database provider and should include information about how estimates have been made.
6) Individuals and contexts are described by attributes. Each database can choose which attributes to provide.
7) Attribute definitions are embedded in the IDS by the attribute Type. A Metadata Registry will be maintained so that common attribute types can be re-used by various archives, but each data provider can define (and register) new attribute types as necessary.
8) Each record entails only one attribute. This approach is known as the Entity Attribute Data Model (EAV) or object-attribute-value model and was already introduced in the 1970s (Stead at al, 1982).

4.3 Data Model

4.3.1 Tables

The IDS consists of five files (or “tables” in database terminology):

INDIVIDUAL consists of attributes belonging to a person (name, sex, wealth, literacy, etc.) and events (birth, marriage, migration, death, etc.). Every item of information about an indi-

vidual is recorded as a separate row in this table. Each row has an attribute type, keys linking to an individual, and a timestamp. Rows in this table may be time-constant attributes (sex, date of birth), time-varying attributes (marital status, occupation), or events that mark changes in attributes (marriage, retirement). The attribute type will distinguish between a marriage certificate (which records the date that a subject's marital status changed from "single" to "married") from the marital status "married" recorded in a census (which means that the subject became married some time before the date of the census).

INDIV_INDIV characterizes relationship between persons. This table will record relationships between two individuals. These relationships may be biological (parent-child), social (husband-wife, godparent-godchild), or economic (master-apprentice, owner-renter). Relationships will be timestamped, when appropriate (e.g. date of marriage).

CONTEXT describes places or environments that affect one or many persons, such as a household, house, geographic location, school, business firm, or organization. Contexts are sets of characteristics shared by groups. Household, for example, implies that a group of individuals shares a common living area, eats together, and pools resources. Contexts may also be places (buildings, geographic coordinates, villages, districts), organizations (business firms), or kinship groups (clans). Like the Individual attribute table, contexts are described by attribute types and timestamps. Contexts may also be layered, and each context may include a link to a higher level of context in which it is nested.

INDIV_CONTEXT associates an individual with a context at a moment or during a period of time. Datestamped links between individuals and contexts are recorded in this table.

METADATA Attribute types will be recorded in a central metadata registry. This will encourage standardization, but it also allows databases to add attribute types that are tailored to their needs. For example, "marriage" will be used by many databases, but some databases will have "publication of marriage banns" and "marriage contract signed."

4.3.2 Individual Data

4.3.2.1 Table Individual

The table INDIVIDUAL contains all attributes that characterize an individual. This table has the following (basic) structure (see also table 1 with some examples of records):

Id — Primary key

Id_D — Identifier of the database or parts of the database from which the data are extracted. This code is especially needed to differentiate between databases in case tables from different databases are merged.

Id_I — Identifying number of each individual in the database. This presupposes that all the identifying work of linking individuals has been done by the database itself.

Source — Specification of the source. We include a field for the source, because an attribute may be reported more than once in different documents within a single database.

Type — Type of attribute (including events that are a subcategory of attributes).
Attribute types are explained in the metadata table. The following examples illustrate attribute types, starting with common ones and ending with more specific attributes belonging to only one database:

- Last name
- Date of Birth
- Location of Birth
- County of location of Birth
- Date of Baptism
- Date of Death
- Date of Marriage
- Location of Marriage
- If the sequence of marriage can be distinguished:
  Date of First marriage
  Date of Second marriage, etc.
- Start observation
- End observation
- Migration move
- Reason for sampling
- Dutch Personal Income Tax (period 1860-1880)

– Number of food distribution Card during First World War

Value The value of the attribute. Many attributes have values, such as "male" and "female" for the attribute "sex." For events (e.g. birth, death), this value usually will be left empty, because the time stamp shows when the event occurred.

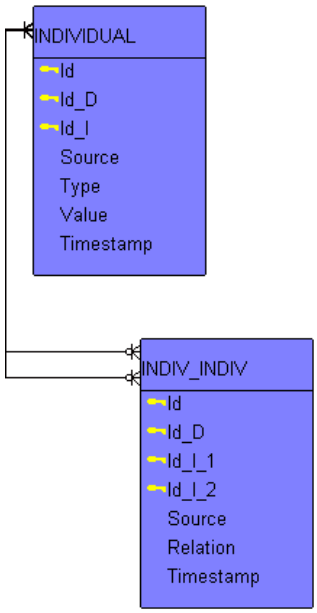
Timestamp A time stamp for the moment or period in time that the attribute is valid (see section 4.4).

Table 1: Records in the table INDIVIDUAL (excluding timestamp variables).

Id	Id_D	Id_I	Source	Type	Value
1	DDB	1	Population Register	Last name	Johansson
2	DDB	1	Population Register	First name	Christiaan
3	DDB	1	Population Register	Date of birth	<time stamp>
4	DDB	1	Population Register	Location of birth	Umeå
4	DDB	1	Population Register	County of location of birth	Västerbotten
6	DDB	1	Population Register	Date of death	<time stamp>
7	DDB	1	Marriage certificate	Date of first marriage	<time stamp>
8	DDB	1	Population Register	Start observation	<time stamp>
9	DDB	1	Population Register	End observation	<time stamp>
10	DDB	1	Income tax register	Occupational title (original)	Timmerman
11	DDB	1	Income tax register	Occupational title (English)	Carpenter
12	DDB	1	Income tax register	Occupational title (HISCO basic)	95410
13	DDB	1	Population Register	Civil status	Married
14	DDB	1	Population Register	Sex	Male
15	DDB	1	Income tax register	Income in Kroner	300
16	DDB	1	Income tax register	Income in dollars 1948	25

4.3.2.2 Table INDIV_INDIV

Figure 3: ERD-diagram tables of individual data.



Explanation: The relations are described by way of so-called Entity_Relationship Diagramming. Here: Every individual may have one or more relationships with other individuals, but every relationship must refer to two individuals in the INDIV_INDIV table (see Beaumont 2007, for more information about Entity Relationship Diagramming).

The table INDIV_INDIV shows how individuals are related to each other. See figure 3 for a presentation of how the INDIVIDUAL and INDIV_INDIV are used, see table 2 for an example of records. This table has the following structure:

Id — Primary key

Id_D — Identifier of the database or parts of the database from which the data are extracted.

Id_I_1 — Identifying number of the first individual in the relationship, referring to Id_I in the first layer

Id_I_2 — Identifying number of the second individual in the relationship, referring to Id_I in the first layer

Source — Specification of the source

Relation — Type of relationship The first part of the relationship refers to Id_I_1, the second part to Id_I_2, for example:

- Father and child
- Bride and groom
- Householder and maid
- etc

Timestamp — A time stamp for the moment or period in time that the relationship is valid (see section 4.4). Some relationships are independent of time, like 'father and child' or 'brother and sister.' The timestamp may be left empty in those cases. The data producer will be responsible for resolving inconsistencies in relationships before the data is transferred into the IDS, but standard programs for detecting inconsistencies may be developed.

Table 2: Records in the table INDIV_INDIV (excluding timestamp variables).

Id	Id_D	Id_I_1	Id_I_2	Source	Relation
1	HSN	1	21	Birth certificate	Mother and child
2	HSN	2	1	Population Register	Husband and wife
3	HSN	1	22	Birth certificate	Mother and child
4	HSN	1	23	Birth certificate	Mother and child
5	HSN	2	21	Population Register	Father and child
6	HSN	2	22	Marriage certificate	Father and child
7	HSN	2	23	Population Register	Father and child
8	HSN	2	234	Population Register	Householder and maid
9	HSN	2	8493	Population Register	Master and apprentice
10	HSN	823	824	Population Register	Brother and sister
11	HSN	823	825	Population Register	Brother and sister
12	HSN	824	825	Population Register	Sister and sister

4.3.3 Context data

4.3.3.1 Table CONTEXT

The CONTEXT table contains information about shared environments, such as households and regions. Each context is assigned a unique identifier by ID_C. Like the INDIVIDUAL table, each row in the CONTEXT table describes an attribute of a context. Constructed attributes (like household size or household type) may be provided by the database as a service to users, but the IDS also allows these attributes to be constructed dynamically by data extraction programs.

An individual can live at the same time in different contexts because they are layered, see further section 4.3.3.4 (and examples in table 3). The CONTEXT table is a table with the following (basic) data structure:

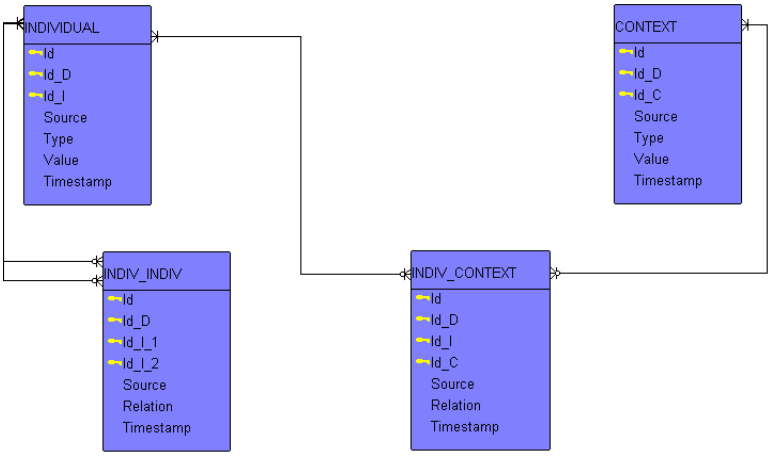
Id	Primary key
Id_D	Identifier of the database or parts of the database from which the data are extracted.
Id_C	Identifying number of the context
Source	Specification of the source
Type	Type of attribute of the context – Name – Layer – Housenumber – Streetname – Postal code – Locality – Municipality – Etc.
Value	The value of the attribute
Timestamp	A time stamp for the moment or period in time that the attribute is valid, see section 4.4. If no timestamp is given in the table CONTEXT the timestamp in the table INDIV_CONTEXT is supposed to cover fully the specific context.

4.3.3.2 Table INDIV_CONTEXT

This table places individuals into contexts. Figure 4 shows how individuals are linked to contexts and to other individuals sharing a common context, see table 3 for an example of records.

Id	Primary key
Id_D	Identifier of the database or parts of the database from which the data are extracted.
Id_I	Identifying number of an individual
Id_C	Identifying number of a context
Source	Specification of the source
Relation	The type of the relationship between individual and context (a value will not always be needed). – Legal membership – Factual membership – Type membership unclear – Head of household (according to source) – Head of household (constructed by rule ##) – Co-resident – Lodger – etc.
Timestamp	A time stamp for the moment or period in time that the attribute is valid, see section 4.4.

Figure 4: ERD-diagram of the Intermediate Data Structure.



Explanation: The relations are described by way of so-called Entity_Relationship Diagramming. Here: Every individual may have one or more relationships with other individuals, but every relationship must refer to two individuals in the INDIV_INDIV table (see Beaumont 2007, for more information about Entity Relationship Diagramming).

4.3.3.3 Households

The concept of 'household' is often problematic. 'Household' usually refers to a group who pool income and share consumption (Hammel and Laslett, 1974; Brettell 2003). In some cultures, households have a continuity over time that is independent of the people that inhabit them. In other cultures, households are simply the group that lives together at a moment in time. In these cases, it is often useful to define households by associating each household with a single reference person, who may or may not be the 'head,' such that everyone who lives with the reference person is in the same household. When a source, such as a census, specifies relationships among people in a household, those relationships can be captured in the INDIV_INDIV table.

4.3.3.4 Context Hierarchies

Contexts are often hierarchical or nested. There are several ways to represent context hierarchies in the IDS. For example, consider a database in which addresses are located in neighbourhoods, which are parts of municipalities. We can represent that information in at least three different ways. The differences between these approaches become clearer if we consider a change in an attribute of a higher level context, for example the population of a municipality.[3] For examples of the following options, see the tables in Appendix A.

1) *Characteristics of higher level contexts may be included as attributes of the most basic context*. Thus, variables describing neighbourhood and municipality may be included as attributes of an address. This involves repetition in the database, because the same attributes are given for all the neighbourhoods in a municipality and for all the municipalities. Also when attributes change the whole has to be repeated but no timestamp is needed because this is defined in the table INDIV_CONTEXT.
2) *Each level in the hierarchy may be represented as a separate context with links from every individual to every level of context in the INDIV_CONTEXT table*. Since each neighbourhood and municipality would be identified by its own ID_C, their attributes would be described only once, and information would not be repeated in the CONTEXT table.

[3] Note that when an individual moves from one context to another at the lowest level in the contextual hierarchy, e.g. address, it is always represented by adding a new row/s to the INDIV_CONTEXT table.

However, every individual would have three rows in the INDIV_CONTEXT table: one for neighbourhood, one for municipality, and one for province. All records in the CONTEXT table need a time stamp otherwise the timestamp of the record in INDIV_CONTEXT will define the period. A change in an attribute of a municipality would result in only one new timestamped attribute, which is associated with the ID_C of the municipality.

3) *Each level of context may be treated as an attribute of the level beneath it.* Thus, "municipality-ID" would be considered an attribute type belonging to neighbourhood, and "neighbourhood-ID" would be an attribute type belonging to an address. As in the second option, each neighbourhood and municipality would be identified by its own ID_C, and its attributes would appear only once in the CONTEXT table. A neighbourhood would be linked to its municipality by putting the ID_C of the municipality in the value column of the CONTEXT table for the attribute "municipality-ID." Each individual would be linked to a neighbourhood with one row in the INDIV_CONTEXT table, and individuals would be linked to municipalities through the "municipality-ID" attribute of the neighbourhood.

In the end, we want the attributes of all three levels of the hierarchy (address, neighbourhood, municipality) to appear in separate columns on every individual record in the rectangular dataset that is used for analysis. Descriptions of higher level contexts are always repeated in the rectangularized file, even if they are not repeated in the IDS.

The trade-off in these three approaches is between repeating information and using more complex programming techniques. Option 1 requires more programming when data is transferred to the IDS, but it simplifies data extraction programs. Options 2 and 3 will result in the most parsimonious IDS tables, but they require more programming to extract information about higher levels in the context hierarchy. Data producers can choose which approach best fits their database, and researchers (data consumers) will make their preferences known in their own ways.

4.3.4 METADATA table

It is important to notice that the variable Type already includes a brief description of the meaning of the attribute. The METADATA table provides a more complete explanation. See figure 5 for the structure of the IDS, including the METADATA table. The METADATA table consists of five fields, the first four form the key to the other tables.

Id	Primary key
Id_D	Identifier of the database or parts of the database from which the

data are extracted. The name 'STANDARD' is reserved for meta-data accepted by the community of researchers for general use, see below.

Type_T Identifier of the table or timestamp concerning the specific meta-data. (all four data tables include a column identifying a type of attribute or relation, and there are three kinds of information about dates on each timestamp, see section 4.4).

- INDIVIDUAL_type
- INDIV_INDIV_relation
- CONTEXT_type
- INDIV_CONTEXT_relation
- TIMESTAMP_date
- TIMESTAMP_estimate
- TIMESTAMP_missing

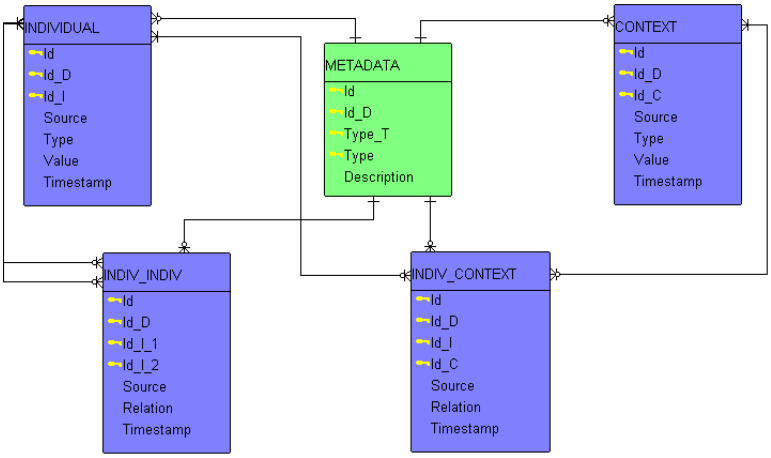
Type Type of attribute, relation or timestamp

Description Memo-field with an explanation of the meaning and use of this type of data (including for example a further description of the relevant sources).

Table 3: Records in the table METADATA

Id	Id_D	Type_T	Type	Explanation
1	STANDARD	INDIVIDUAL	DEATH	Date of occurrence of death
2	HSN	INDIVIDUAL	DEATH	Standard, three sources which we used in the following preference : 1 civil certificate, 2 population register, 3 Red Cross. We use "Red cross" as source when dates are estimated on the basis of circumstantial information but must be considered quite accurate, e.g. the date of death in German termination camps like Sobibor which was estimated on the base of date of deportation from the Netherlands. Civil certificates are only used for persons on which the HSN is based, so-called Research Persons (for more explanation see field SAMPLE)
3	HSN	INDIVIDUAL	DEATH_m	Period of death estimated on the basis of evidence from marriage certificates

Figure 5: ERD-diagram of the Intermediate Data Structure including the metadata table



Explanation: The relations are described by way of so-called Entity_Relationship Diagramming. Here: Every individual may have one or more relationships with other individuals, but every relationship must refer to two individuals in the INDIV_INDIV table (see Beaumont 2007, for more information about Entity Relationship Diagramming).

The value STANDARD in the field ID_D is reserved to distinguish standard definitions of variables from more database specific ones. The STANDARD meaning of an attribute will be specified by the community of researchers, and database-administrators must follow those guidelines, if they use a standard TYPE. Databases will add rows with their own ID_D for each standard TYPE, which they may also use to describe how an attribute is derived from the sources available to them. Thus, a TYPE will have only one row with ID_D=STANDARD, showing the community's specification of this attribute, but it may have many rows explaining if and how various databases implemented that type. Table 3 gives an example of three records in the metadata-registry.

4.4 Time Stamp

Time is defined by way of the Gregorian calendar. We make a distinction between dates and periods. If the reference is an exact date (e.g. a birth date), it is not necessary to define a period. When there is with some degree of fuzziness about a date, we include the period in which the date is situated.

In principle databases will provide estimates of dates in case of missing values. They will describe how they have estimated their dates in the metadata_table by providing an explanation for the Date_estimate_type. Each Time Stamp consists of the following elements (or variables):

Date_Type — Type of each date
- Date of occurrence
- Declared date
- Date based on some kind of estimation
- Etc.

Date_Estimate_Type — Type of estimation of date
- Exact date
- Exact month and year
- Exact year
- Middling of period
- Period of birth based on age and date of source
  For example:
  If you know an age of 25 in the end of the month February 1860, then you know that the person is born between 1st of March 1834 and the 28th of February 1835.
- Etc.

An exact date consists of five variables:

Day	Day number
Month	Month number
Year	Year number
Hour	Hour (0 to 23 hours)
Minute	Minutes (0 to 59 minutes)

A period is defined by six variables

Start_day	Start day number
Start_month	Start month number
Start_year	Start year number
End_day	End day number
End_month	End month number
End_year	End year number

Missing — This field explains why a date or part of a date is missing (Mandemakers and Dillon 2004)

- Unavailable (in the source)
- Unreadable (in the source)
- Anonymized (by the database)
- Private (not available for reasons of privacy, not included in the database)
- Time invariant (e.g. sex)
- Unknown (Unknown in the database why a value is unknown)

Day, month, and year are included as separate columns, rather than relying on the built-in date formats used by various software packages, to avoid incompatibilities between systems. The values of Date_Type, Date_estimate_type and Missing may be further explained in the meta-data registry. A time stamp can be developed much further, including atomic precision but for historical databases a precision in minutes seems to be sufficient (compare J. Benzler & S. Clark 2005).

5. Summary and Perspective

We believe that longitudinal historical databases are extremely rich and valuable resources, which should be much more widely used. But we also understand that they are difficult to use. Every kind of analysis with these data requires some programming, and some of the data construction tasks are complicated. Moreover, problems like selection and informative censoring are subtle and not obvious to the inexperienced. If we really want to expand the use of these databases, we will need to find new ways to reach out to a broader community of researchers. The Intermediate Data Structure approach has a number of important benefits.

1) *It is open, scalable, and extendable*. Any database can transfer its data to the IDS, and the metadata registry will be extended to accommodate new types of data as they become available. New types of analysis can be introduced by adding new extraction modules.
2) *Database managers will decide what data they provide and how their data can be used*. Data producers can transfer data to the IDS in stages. Attributes that require minimal programming can be issued first, and new versions of the database can be created as more difficult data management tasks are solved. Databases that include confidential information can withhold identifiers that would disclose individual identities. For example, databases that have complex censoring structures can develop attribute types that limit the ways that their data are used. Since extraction programs will

require specific attribute types, data providers can be sure that only appropriate data management procedures will be applied to their data.

3) *Extraction programs will be re-usable and transparent*. Anyone can contribute an extraction module, and all extraction modules will operate on every dataset with the required data. Extraction programs will also be open to scrutiny by the research community. Methodologies can be examined, discussed, and tested, and research results will be reproducible.

Data producers often express concerns that inexperienced researchers will misuse and misinterpret their data. They ask: Is the risk of facilitating bad science greater than the advantages of encouraging wider use? This problem has limited the dissemination of data, and it has imposed significant costs on both data producers and researchers. We think that the IDS will alleviate this problem by providing community-based structures that help researchers use data correctly.

We also believe that the IDS must be accompanied by renewed efforts to teach new generations the fundamental insights of historical demography. The central problem is that students interested in historical demography and faculty with the skills to teach them are too widely dispersed, so it is hard to keep a course on methods of historical demography in a regular curriculum. The only way to assemble students is to cooperate across institutions and national borders. There have been efforts in this direction, but they have been sporadic. The historical demography societies of France, Spain, and Italy offered a summer school, but it has not been repeated recently. Courses have also been offered at Lund and the Max Planck Institute. The latest effort of this sort is a four-week summer course in historical demography under the auspices of ICPSR with funding from the U.S. National Institutes of Health, which was held in 2006 and 2007. The IDS will enhance the value of these courses, because it will be a gateway to longitudinal data about past societies all over the world.

References

Alter, G. (1988): Family and the Female Life Course: The Women of Verviers, Belgium, 1849-1880. University of Wisconsin Press

Alter, G./Gutmann, M.P. (1999): Casting Spells. Database Concepts for Event-History Analysis. In: Historical Methods 32 (4). 165-176

Alter, G./Campbell, C./Derosas, R. (2006). Household context and the timing of first marriage in Eurasian comparative. Amsterdam: Paper ESSHC 2006.

Beaumont, R. (2007): An Introduction to Entity Relationship Diagrams (ERDs), version 5, http://www.visualcplusdotnet.com/erddatabasemodeling.pdf

Bengtsson, T./Campbell, C./Lee, J.Z. (2004): Life under Pressure. Mortality and Living Standards in Europe and Asia, 1700-1900. Cambridge Massachusetts/London England: The MIT Press

Benzler, J./Clark, S. (2005): Toward a Unified Timestamp with explicit precision. In: Demographic Research 12. 107-140

Brettell, C. (2003): Anthropology and Migration. Essays in Transnationalism. Ethnicity and Identity. Walnut Creek: Alatamira Press

Dillon, L./Roberts, E. (2006): Introduction: Longitudinal and Cross-sectional Historical data: Intersections and Opportunities. In: History and Computing 14. 1-8

Elder, G.J. Jr./Kirkpatrick Johnson, M./Crosnoe, R. (2003): The Emergence and Development of Life Course Theory. In: Mortimer, J.T./Shanahan, M.J. (Eds.) Handbook of the Life Course. New York: Plenum. 3-19

Hammel, E.A./Laslett, P. (1974): Comparing Household Structure over Time and Between Cultures. In: Comparative Studies in Society and History 16 (1). 73-109

Hareven, T.K. (1982): Family time and industrial time: the relationship between the family and work in a New England industrial community. Cambridge: Cambridge University Press.

Kelly Hall, P./McCaa, R./Thorvaldsen, G. (Eds.) (2000): Handbook of International Historical Microdata for Population Research. Minneapolis: Minnesota Population Center

Kok, J./Mandemakers, K./Wals, H. (2005): City nomads: Changing Residence as a Coping Strategy, Amsterdam, 1890-1940. In: Social Science History 29 (1). 15-44

Mandemakers, K./Dillon, L. (2004): Best practices with large databases on historical populations. In: Historical Methods 37 (1). 34-38

Mandemakers, K. (2008): Working with Intermediate Structures as a condition for easy building of datasets from historical longitudinal databases. Third version. Paper presented at seminar DDB Umea (2007), European Social Science History Conference (2008) and the RC33 conference in Naples, September 2008.

Ruggles, S./Sobek, M./Alexander, T./Fitch, C. A./Goeken, R./Kelly Hall, P./King, M./Ronnander, C. (2008): Integrated Public Use Microdata Series: Version 4.0 [Machine-readable database]. Minneapolis, MN: Minnesota Population Center.

Settersten, R.A. (Ed.) (2002): Invitation to the life course: toward new understandings of later life. Amityville: Baywood Publishing.

Smith, D. S. (1973): Parental Power and Marriage Patterns: An Analysis of Historical Trends in Hingham, Massachusetts. In: Journal of Marriage and the Family 35. 419-428

Stead, W.W./Hammond, W.E./Straube, M.J. (1982): A chartless record – Is it adequate? In: Proceedings of the Annual Symposium on Computer Application in Medical Care 11-02. 89–94

Appendix A.: Examples of the different options fore context tables

Table 4: Example of the CONTEX table, built with option 1 (see section 4.3.3.4)

CONTEXT table

ID	Id_D	Id_C	Type	Value
1	Utah	115023	Streetname	Mainstreet
2	Utah	115023	Streetnumber	12
3	Utah	115023	Longitude	233,838
4	Utah	115023	Latitude	193,933
5	Utah	115023	Direction	112,23
6	Utah	115023	Locality_name	Salt Lake Harbour
7	Utah	115023	Number _inhab	230
8	Utah	115023	Long_Centroid	233,838
9	Utah	115023	Latit_Centroid	193,933
10	Utah	115023	Municipality Name	Salt Lake City
11	Utah	115023	Number _inhab	23455
12	Utah	115023	Long_Centroid	233,921
13	Utah	115023	Latit_Centroid	193,888
14	Utah	115029	Streetname	Smallstreet
15	Utah	115029	Streetnumber	212
16	Utah	115029	Longitude	233,847
17	Utah	115029	Latitude	193,899
18	Utah	115029	Direction	247,77
19	Utah	115029	Locality_name	Salt Lake Harbour

ID	Id_D	Id_C	Type	Value
20	Utah	115029	Number _inhab	230
21	Utah	115029	Long_Centroid	233,838
22	Utah	115029	Latit_Centroid	193,933
23	Utah	115029	Municipality Name	Salt Lake City
24	Utah	115029	Number _inhab	23455
25	Utah	115029	Long_Centroid	233,921
26	Utah	115029	Latit_Centroid	193,888

INDIVIDUAL table

ID	Id_D	Id_I	Type	Value
1	Utah	1001	First name	Fred
2	Utah	1001	Last name	Jones
3	Utah	2009	First name	Samantha
4	Utah	2009	Last name	Smith

INDIV_CONTEXT table

ID	Id_D	Id_I	Id_C	Relation	Start_day	Start_month	Start_year	End_ day	End_month	End_year
1	Utah	1001	115023	Address	21	2	1879	2	6	1880
2	Utah	2009	115029	Address	15	8	1879	5	8	1882

Table 5: Example of the CONTEX table, built with option 2 (see section 4.3.3.4).

CONTEXT table

ID	Id_D	Id_C	Type	Value	Start_day	Start_month	Start_year	End_day	End_month	End_year
1	Utah	115023	Level_name	Address						
2	Utah	115023	Streetname	Mainstreet						
3	Utah	115023	Streetnumber	12						
4	Utah	115023	Longitude	233,838						
5	Utah	115023	Latitude	193,933						
6	Utah	115023	Direction	112,23						
7	Utah	115029	Level_name	Address						
8	Utah	115029	Streetname	Smallstreet						
9	Utah	115029	Streetnumber	212						
10	Utah	115029	Longitude	233,847						
11	Utah	115029	Latitude	193,899						
12	Utah	115029	Direction	247,77						
13	Utah	9022	Level_name	Locality	1	1	1850	31	12	1910
14	Utah	9022	Locality_name	Salt Lake Harbour	1	1	1850	31	12	1910
15	Utah	9022	Number _inhab	230	1	1	1850	31	12	1910
16	Utah	9022	Long_Centroid	233,838	1	1	1850	31	12	1910
17	Utah	9022	Latit_Centroid	193,933	1	1	1850	31	12	1910
18	Utah	10345	Level_name	Municipality	1	1	1850	31	12	1910
19	Utah	10345	Name	Salt Lake City	1	1	1850	31	12	1910
20	Utah	10345	Number _inhab	23455	1	1	1879	31	12	1879
21	Utah	10345	Number _inhab	23655	1	1	1880	31	12	1880
22	Utah	10345	Number _inhab	23867	1	1	1881	31	12	1881
23	Utah	10345	Number _inhab	23841	1	1	1882	31	12	1882
24	Utah	10345	Long_Centroid	233,921	1	1	1850	31	12	1910

ID	Id_D	Id_C	Type	Value	Start_day	Start_month	Start_year	End_day	End_month	End_year
25	Utah	10345	Latit_Centroid	193,888	1	1	1850	31	12	1910

INDIVIDUAL table

ID	Id_D	Id_I	Type	Value
1	Utah	1001	First name	Fred
2	Utah	1001	Last name	Jones
3	Utah	2009	First name	Samantha
4	Utah	2009	Last name	Smith

INDIV_CONTEXT table

ID	Id_D	Id_I	Id_C	Relation						
1	Utah	1001	115023	Address	21	2	1879	2	6	1880
2	Utah	1001	9022	Locality	21	2	1879	2	6	1880
3	Utah	1001	10345	Municipality	21	2	1879	2	6	1880
4	Utah	1001	115029	Address	3	6	1880	30	11	1882
5	Utah	1001	9022	Locality	3	6	1880	30	11	1882
6	Utah	1001	10345	Municipality	3	6	1880	30	11	1882
7	Utah	2009	115029	Address	15	8	1879	5	8	1882
8	Utah	2009	9022	Locality	15	8	1879	5	8	1882
9	Utah	2009	10345	Municipality	15	8	1879	5	8	1882

Table 6: Example of the CONTEX table, built with option 3 (see section 4.3.3.4).

CONTEXT table

ID	Id_D	Id_C	Type	Value	Start_day	Start_month	Start_year	End_day	End_month	End_year
1	Utah	115023	Level_name	Address						
2	Utah	115023	Streetname	Mainstreet						
3	Utah	115023	Streetnumber	12						
4	Utah	115023	Longitude	233,838						
5	Utah	115023	Latitude	193,933						
6	Utah	115023	Direction	112,23						
7	Utah	115023	Locality_id	9022						
8	Utah	115029	Level_name	Address						
9	Utah	115029	Streetname	Smallstreet						
10	Utah	115029	Streetnumber	212						
11	Utah	115029	Longitude	233,847						
12	Utah	115029	Latitude	193,899						
13	Utah	115029	Direction	247,77						
14	Utah	115029	Locality_id	9022						
15	Utah	9022	Level_name	Locality	1	1	1850	31	12	1910
16	Utah	9022	Locality_name	Salt Lake Harbour	1	1	1850	31	12	1910
17	Utah	9022	Number _inhab	230	1	1	1850	31	12	1910
18	Utah	9022	Long_Centroid	233,838	1	1	1850	31	12	1910
19	Utah	9022	Latit_Centroid	193,933	1	1	1850	31	12	1910
20	Utah	9022	Municipality_id	10345	1	1	1850	31	12	1910
21	Utah	10345	Level_name	Municipality	1	1	1850	31	12	1910
22	Utah	10345	Name	Salt Lake City	1	1	1850	31	12	1910
23	Utah	10345	Number _inhab	23455	1	1	1879	31	12	1879
24	Utah	10345	Number _inhab	23655	1	1	1880	31	12	1880

ID	Id_D	Id_C	Type	Value	Start_day	Start_month	Start_year	End_day	End_month	End_year
25	Utah	10345	Number _inhab	23867	1	1	1881	31	12	1881
26	Utah	10345	Number _inhab	23841	1	1	1882	31	12	1882
27	Utah	10345	Long_Centroid	233,921	1	1	1850	31	12	1910
28	Utah	10345	Latit_Centroid	193,888	1	1	1850	31	12	1910

INDIVIDUAL table

ID	Id_D	Id_I	Type	Value
1	Utah	1001	First name	Fred
2	Utah	1001	Last name	Jones
3	Utah	2009	First name	Samantha
4	Utah	2009	Last name	Smith

INDIV_CONTEXT table

ID	Id_D	Id_I	Id_C	Relation						
1	Utah	1001	115023	Address	21	2	1879	2	6	1880
2	Utah	1001	115029	Address	3	6	1880	30	11	1882
3	Utah	2009	115029	Address	15	8	1879	5	8	1882

Authors

George Alter is Professor of History and Associate Director of the Inter-university Consortium for Social and Political Research at the University of Michigan.

Research Fields: Historical demography, population studies, economic history.

Selected Publications:
Effects of Inheritance and Environment on the Heights of Brothers in Nineteenth-century Belgium, In: Human Nature (with Michel Oris; 2008)
Widowhood, Family Size, and Post-Reproductive Mortality: A Comparative Analysis of Three Populations in Nineteenth Century Europe, In: Demography (with Martin Dribe and Frans van Poppel; 2007)
When Protoindustry Collapsed: Fertility and the Demographic Regime in Rural Eastern Belgium During the Industrial Revolution, In: Historical Social Research (with Michel Oris and Muriel Neven; 2007)
Childhood Conditions, Migration, and Mortality: Migrants and Natives in Nineteenth-century Cities, In: Social Biology (with Michel Oris; 2005)
Le modèle européen de marriage dans une perspective non européene, In: Paul Servais and George Alter (eds.): *Le Mariage dans l'Est de la Wallonie*, Louvain-la-Neuve: Academia-Bruylant (2005);
Height, Frailty, and the Standard of Living: Modeling the Effects of Diet and Disease on Declining Mortality and Increasing Height, In: Population Studies (2004)

Contact Data: altergc@umich.edu

Kees Mandemakers is Senior Research Fellow at the International Institute for Social History (IISH), heading the Historical Sample of the Netherlands (HSN) and Professor of Large Historical Databases at the Faculty of History and Arts, Erasmus University Rotterdam.

Research fields: Methodology of large historical databases, Historical Demography, Social stratification and mobility, Social history of education.

Selected publications:
Time Trends in Social Class Mortality Differentials in the Netherlands 1820-1920: An Assessment based on Indirect Estimation Techniques', In: Social Science History (with Frans van Poppel and Roel Jennissen; 2009)

Marriage Timing over the Generations, In: Human Nature (with Frans van Poppel and Christiaan Monden; 2008)
Building Life Course Datasets from Population Registers by the Historical Sample of the Netherlands (HSN), In: History and Computing (2006)
City Nomads: Changing Residence as a Coping Strategy, Amsterdam, 1890-1940, In: Social Science History (with Jan Kok and Henk Wals; 2005)
Best Practices with Large Databases on Historical Populations, In: Historical Methods *(with Lisa Dillon; 2004)*

Contact Data: kma@iisg.nl

Myron Gutmann is Director of the Inter-university Consortium for Political and Social Research and Professor of History and Information at the University of Michigan, Ann Arbor, Michigan

Research Fields: Historical Demography of Europe and the United States, Methods for effective Archiving and Preservation of Research Data, especially protection of confidentiality of Human Subjecs

Selected publications:
Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data, Washington: National Academy Press (edited with Paul Stern, 2007)
Social Science: Computational Social Science, In: Science (with D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne, 2009)
Building Partnerships Among Social Science Re-searchers, Institution-based Repositories and Domain Specific Data Archives, In: *OCLC Systems & Services: International Digital Library Perspectives* (with Ann Green, 2007)

*Contact Data:*gutmann@umich.edu