The Demographic Database — History of Technical and Methodological Achievements

By Pär Vikström, Maria Larsson, Elisabeth Engberg and Sören Edvinsson

To cite this article: Vikström, P., Larsson, M., Engberg, E., & Edvinsson, S. (2023). The Demographic Database — History of Technical and Methodological Achievements. *Historical Life Course Studies, 13*, 89–102. https://doi.org/10.51964/hlcs12163

HISTORICAL LIFE COURSE STUDIES

Content, Design and Structure of Major Databases with Historical Longitudinal Population Data

VOLUME 13, SPECIAL ISSUE 5

GUEST EDITORS
George Alter
Kees Mandemakers
Hélène Vézina

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 hlcs.nl.

hislives@kuleuven.be

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

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

HISTORICAL LIFE COURSE STUDIES
VOLUME 13 (2023), published 30-03-2023

The Demographic Database — History of Technical and Methodological Achievements

Pär Vikström

Maria Larsson

Elisabeth Engberg

Sören Edvinsson

Centre for Demographic and Ageing Research, Umeå University, Sweden

ABSTRACT

The Demographic Data Base (DDB) at the Centre for Demographic and Ageing Research (CEDAR) at Umeå University has since the 1970s been building longitudinal population databases and disseminating data for research. The databases were built to serve as national research infrastructures, useful for addressing an indefinite number of research questions within a broad range of scientific fields, and open to all academic researchers who wanted to use the data. A countless number of customized datasets have been prepared and distributed to researchers in Sweden and abroad and to date, the research has resulted in more than a thousand published scientific reports, books, and articles within a broad range of academic fields. This article will focus on the development of techniques and methods used to store and structure the data at DDB from the beginning in 1973 until today. This includes digitization methods, database design and methods for linkage. The different systems developed for implementing these methods are also described and to some extent, the hardware used.

Keywords: Database, Linkage, RDBMS, Digitization, Church registers

e-ISSN: 2352-6343
DOI article: https://doi.org/10.51964/hlcs12163

1 INTRODUCTION

Established in the infancy of historical demography, as Lionel Kesztenbaum in a recent article defines the period that ended in the early 1970s, the Demographic Data Base (DDB) at Umeå University has experienced remarkable technical and methodological development (Kesztenbaum, 2021). Manual excerpts, punch cards, and magnetic tapes are long gone, replaced by automated and semi-automated processes, custom-made software, and modern server technology (Edvinsson & Engberg, 2020). This article gives a brief description of the different technical approaches and methods used by the DDB to design, build, structure, and store large population databases since the start in 1973. In the first part, particular attention is given to the development of different structures to store data, and to digitization systems and hardware. The second part focuses on the development of different linkage methods and processes, from manual and semi-manual methods to fully automated processes.

Today, DDB owns and administers three main research databases: a) POPUM, with individual-level data from Swedish parishes in different areas, see Figure 1, covering the period 1680–1900; b) POPLINK, with similar data but covering a longer time span, until around 1950; and c) TABVERK, with aggregate statistics from all Swedish parishes for the period 1749–1859. POPUM and POPLINK are some of the most detailed historical databases in the world when it comes to the wealth of information per individual. The production database KBGRUNDS stores all digitized information from these registers, see Table 1 with an overview of the data as released on 2021-11-15. The latest version of POPUM and POPLINK, version 6.4.1, based on this release of KBGRUNDS, was released on 2021-12-16, see Table 2.

Already from the start, the databases at the DDB were built to serve as research infrastructures, useful for addressing an indefinite number of research questions within a broad range of scientific fields, and open to all academic researchers who want to use the data (Edvinsson & Engberg, 2020). The databases are based on the Swedish parish registers, which from 1680 until 1990 served as the system of official registration. Hence, the registers do not only include the majority population belonging to the Lutheran national state-church system, but also the part of the population belonging to other faiths and denominations. This means that the Swedish parish registers have a nearly unsurpassed coverage of the entire population. The records include births, marriages, and deaths, as well as family-based continuous registers covering the entire population, and their long time spans, from the late 17th century and forward, offer virtually unparalleled possibilities for longitudinal studies (Nilsdotter Jeub, 1993).

Although the technical and methodological changes have been immense, the principles guiding the basic demands of the longitudinal population databases have remained fairly unchanged (Johansson & Åkerman, 1973; Vikström, Edvinsson, & Brändström, 2006):

1. The database shall be *true to the source*. It must be possible to trace all records back to the original source for verification.
2. The database shall be *complete*; that is, all relevant information in the original source shall be included in the data collection.
3. The data collection shall be *open*, which means that the database shall be built in a way that allows the inclusion of new data, in time as well as in space.
4. The database shall be *coherent* and *consistent*: data entry shall be performed according to similar rules and principles, for maximum comparability and coordination.
5. Data entry, processing, and storage shall be performed in an *efficient* way.
6. All processing of data shall be *research-oriented*, allowing for micro-historic research as well as large-scale cohort studies.

Similar principles have later also been formulated by others, for example by Mandemakers and Dillon (2004) as best practice in the field of building longitudinal historical databases for research (Edvinsson & Engberg, 2020).

Figure 1 *Geographical coverage of POPUM and POPLINK version 6.4.1*

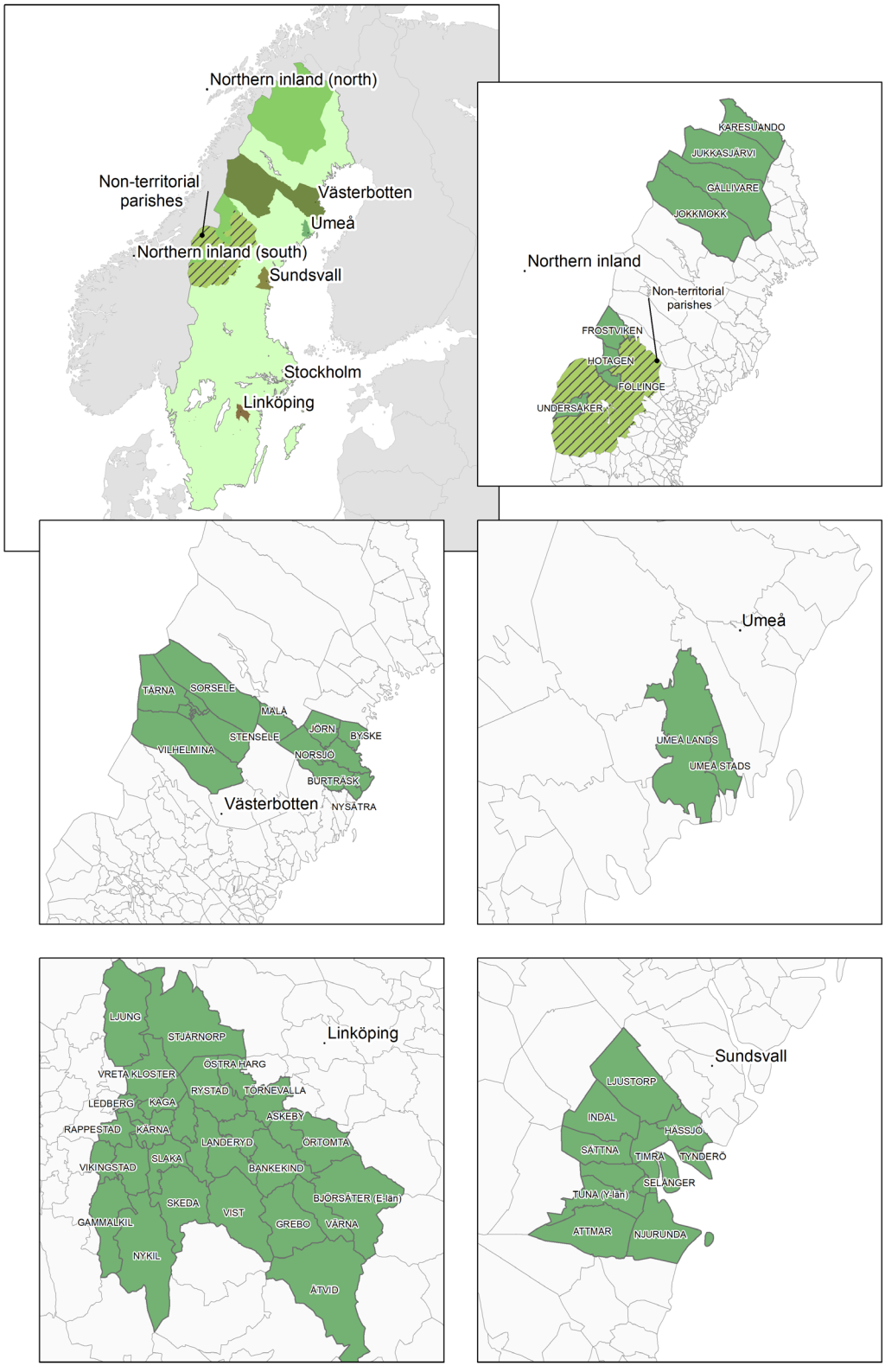


Table 1 *Number of source records in KBGRUNDS 2021-11-15*

Source (register)/ Region	Linköping	Northern inland	Sundsvall	Umeå	Västerbotten	TOTAL
Death and burial	167.621	30.898	50.235	20.452	150.793	419.999
Birth and baptism	229.745	54.856	105.623	40.149	342.288	772.661
Longitudinal parish	1.605.805	217.843	537.930	293.599	1.831.942	4.487.119
In-migration	202.147	7.048	83.358	67.254	161.383	521.190
Banns and marriage	35.243	13.438	25.453	12.505	76.928	163.567
Out-migration	225.284	5.461	72.232	58.061	186.516	547.554
TOTAL	2.465.845	329.544	874.831	492.020	2.749.850	6.912.090

Table 2 *Number of individuals in POPUM and POPLINK version6_4_1 (released 2021-12-16)*

Region	Linköping	Northern inland		Sundsvall	Umeå	Västerbotten		TOTAL	
Sex/Database	POPUM	POPUM	POPLINK	POPUM	POPLINK	POPUM	POPLINK	POPUM	POPLINK
Unknown	5.618	1.033	54	189	9	1.141	1.328	7.981	1.391
Male	325.339	46.073	6.243	88.673	56.142	114.240	198.574	574.325	260.959
Female	334.024	45.063	6.543	83.021	57.777	111.313	191.924	573.421	256.244
TOTAL	664.981	92.169	12.840	171.883	113.928	226.694	391.826	1.155.727	518.594

2 DATABASE STRUCTURE AND DATA ENTRY SYSTEMS

2.1 DATA MANAGEMENT SYSTEMS BEFORE THE ERA OF RELATIONAL DATABASES

In 1973, when the data collection for the DDB-databases began, data entry was essentially a manual process. Between 1973 and 1982, data entry centers were established in six different locations in northern Sweden, funded by provisional contributions from the National Labour Market Board. At the height of operations, more than 100 data entry assistants worked at these centers. Information from the sources was transcribed by hand into forms on printed cards, one card for each entry. These paper cards were then manually linked together by ordering them into bundles, each bundle describing one individual, and the information was digitized using punch cards (Edvinsson & Engberg, 2020). Two technical hubs for this kind of work were set up, one in Umeå and one in Haparanda, where the very first data entry centre was established. After a couple of years the punch cards were abandoned and a commercial digitization system, also called "KEY to DISK", was introduced, and the bundles of cards were digitized on small terminals. This made the process of digitization more efficient. The first system of this kind was a CMC 12 (Communication Machinery Company) upgraded after five years to a CMC 5400.

The Swedish parish registers are ordered into an intricate system, making up something that almost resembles a kind of non-digitalized relational database structure. The base in the system is the catechetical register, a household-based longitudinal parish register which is a distinctive feature of the Swedish registers, providing basic demographic information about the entire population in a parish. Like in a census, families were kept together on the same page. When a child was born, a parent died, or a widow remarried, it was not only noted in the separate event registers, but also in the longitudinal register. Record linkage is facilitated by clever links between said registers. The longitudinal registers include references to the volume and page, where first-hand information about a particular birth, marriage, or death can be found. The event registers are linked in a similar way to the longitudinal registers, creating a comprehensive system of information whereby individuals can be followed over their entire life spans. With this kind of double bookkeeping, it is also possible to reconstruct missing or

lost information. When a longitudinal register volume was completed after five to ten years a new one was established, into which the minister transferred current information from the old volume, of course with valuable links between the old and new registers. For obvious reasons, the longitudinal registers are extremely valuable for the creation of life-course data, making it possible to follow individuals and families over their entire life spans and over generations as long as they remain within the parish borders (Edvinsson & Engberg, 2020).

This comprehensive structure of the sources was at the beginning used as a template to build a database structure consisting of flat files, called Individ. These files served well to store the information, but the flat format was not optimal for data extraction, both in the case of defining a cohort for research but also in extracting the variables required. The Individ-files also had other limitations. Due to the digitization systems used during the first years, before relational databases were introduced, the records had a maximum length of 256 characters, which usually resulted in one record from the source becoming two records in the file. To solve this problem, the Individ-files were converted into a 518-character format, called INDIVSQ, which covered the full record from the source. This conversion was made by the DDB, after the information had been transferred to hardware at the computer centre at Umeå University, UMDAC. Much of the programming and other information processing at the DDB was, during the end of the 1970's and the beginning of the 1980's, performed by using a terminal connection to UMDAC and portable printing terminals, Texas Silent.

Despite the limitations in database structure, a large amount of data was collected and digitized during these early years. Records were transcribed, linked manually and stored in a database management system. The first parish to be digitized was Tuna, a small parish in the industrialized Sundsvall area, which at that time was already an object of interest among historians and social scientists as being one of the fastest growing industrial regions in Sweden (Brändström, 2009). The work continued with records from six other parishes in different Swedish regions, Locknevi, Gullholmen, Trosa, Fleninge, Nedertorneå and Svinnegarn, selected on scientific merits expressed by researchers. Nedertorneå, was for example chosen because of its high level of infant mortality (Brändström, 1984). Together, these seven parishes formed the beginning of the database that later would be named POPUM. The first large region included in POPUM was the industrialized Sundsvall area, a previously agricultural district that in the 19th century became the heart of the sawmill industry in Europe and a center for the Swedish labor movement. In the early 1980s, the Skellefteå area in the north was selected for a large project in genetic epidemiology (Edvinsson & Engberg, 2020). By then, small Luxor ABC802-computers were used for digitization. The diskettes were continuously transferred by mail from the data entry centers to the technical hub in Umeå, copied to the computers at UMDAC and later converted to INDIVSQ-format.

2.2 RELATIONAL DATABASE MANAGEMENT SYSTEMS

But soon, a new and influential general theory of data management would change the situation and solve the problems with the flat files. In 1970 Dr E. F. Codd presented his seminal work "A Relational Model of Data for Large Shared Data Banks" (Codd, 1970), where he presented a theoretical model for relational databases and relational database management systems that would have a large impact upon database modelling and significantly facilitate the development of large databases, including the DDB.

In 1983 DDB decided to start converting all data into a relational database, using the model that E. F. Codd had presented 13 years earlier. The first relational database management system (RDBMS) used by DDB was IBM SQL/DS. With this new system it was possible to use the SQL standard language to select and retrieve data from the database, and to join information from different tables. Making simple retrievals was easy to accomplish.

However, the transition from the old to the new database management system did not take place without difficulties, and it turned out to be a more time-consuming enterprise than originally planned. The first preparations for the conversion were made in 1983, by specifying demands on the new database structure and the practical work started in 1985. Almost all hardware had to be installed at the same time. It would have been better if one of the systems had been installed as a pilot system, making it easier to identify problems and how to solve these, before the second system was installed.

The hardware consisted of a minicomputer system from IBM called 4361 VM/CMS, one in Umeå and one in Haparanda. A somewhat special situation occurred when installing the hardware in Haparanda. A window, and a part of the wall of the building had to be removed, to be able to lift in the minicomputer, as this quite huge machine actually was called.

In this novel technical environment, digitization, programming, and database management were performed from terminals connected directly to the two minicomputers. New software had to be developed for the conversion of INDIVSQ-files to a relational database and a new data entry system also had to be designed and developed to interact with the technical environment. At the same time the staff at DDB also had to be trained to be able to adapt to a completely new way of data processing.

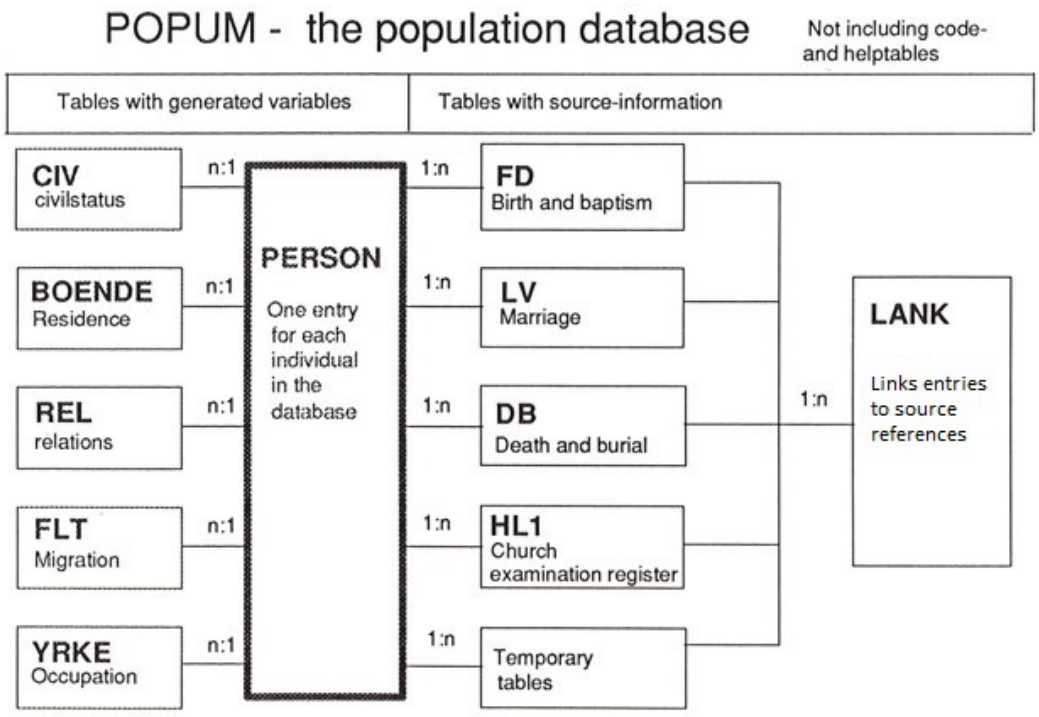
This whole situation caused several problems. One was that the pace of digitization slowed down. Another, even more crucial problem, was that the research using DDB-data also slowed down. For a period of almost three years very few retrievals for researchers were made.

After more than two years, in 1987, the very first version, a beta version, of the relational database was ready for use. It did not have a name until version 2 of the database was published in 1991. Then the database was given the name POPUM, made up from two central concepts, Population and Umeå. An overview of this early relational database is presented in Figure 2.

So, why did the conversion end up taking three years instead of the nine months as originally planned? The answer is simple. It was pioneering work. No one at DDB had made this kind of transition before. However, knowledge and competence accumulated during this process has made later transitions easier.

Although the new relational database management system significantly improved the data production process, there were still technical issues that caused problems and required attention. The most problematic issue was perhaps limitations in disk space. Disks were extremely expensive, and the two minicomputers did not have enough capacity to handle the required amount of data. A lot of time was spent finding solutions to work with large amounts of data from big parishes. The solution was to store data from one part of the parish on magnetic tapes, while dealing with the other part on disk. In 1994 the issue of disk space became less problematic, when the old IBM 4361 system gave way for new IBM RS6000-servers with the IBM dialect of UNIX, AIX. All terminals were substituted with personal computers connected in a network. At the same time, the relational database management system was changed from SQL/DS to Ingres, mostly because SQL/DS did not work with AIX. A new data entry system, RODE, was also developed and used on the personal computers. Files from RODE were transferred from the personal computers over the network to a production database, POPHAP, for subsequent processing of the data, involving coding, standardization, and linkage, before finally being included in POPUM. POPHAP was used until 2005, in the end mostly for linkage purposes.

Figure 2 *POPUM version 2*



2.3 A NEW MILLENNIUM

At the end of the previous millennium the RS6000-servers were getting old and had to be replaced. Replacing them with new RS6000 would have been too expensive, so in 1998 it was decided that they should be replaced with INTEL-based servers from DELL and HP. INGRES was replaced with IBM DB2 in 1999. Around the same time, a new digitization system, REGINA, was developed, primarily for adaption to the new database management system but also to avoid problems with possible millennium bugs.

Over the years it had become evident that not all the requirements on the database structure set in 1983 had been met. This, in combination with a normalization that needed some adjustments and new demands caused by methodological developments in research, made it necessary to conduct a larger review of the database structure. Another issue that had become more and more disturbing over the years, was the use of multiple digitization systems for entering data. The data, and in particular the coding schemes from the different systems were not always harmonized, which caused a lot of extra programming when data was extracted for researchers. These issues were finally solved with version 3 of POPUM, in 2006.

Along with version 3 of POPUM a new database was added to improve the normalization and avoid inconsistency in data, the database KBGRUND. This new database constitutes the basic database at DDB, containing all information that has been digitized, and into which all new data is stored and post processed. All information in KBGRUND is traceable back to the original source, not only to the record but also to the individuals mentioned in the source record, i.e., in the birth register we have, not only the child, but also its mother and father. KBGRUND is a strictly normalized database. POPUM version 3, on the other hand, is a user database containing specific post-processed tables describing certain events and states, like migration and residence, as well as all tables from KBGRUND. POPUM is mainly used for retrieval of subsets of the data for research and therefore a certain degree of denormalization is allowed. But to avoid inconsistencies, the database is only available as read-only. Version 3 of KBGRUND was released at the same time as REGINA in 2000 and version 3 of POPUM was released in 2006.

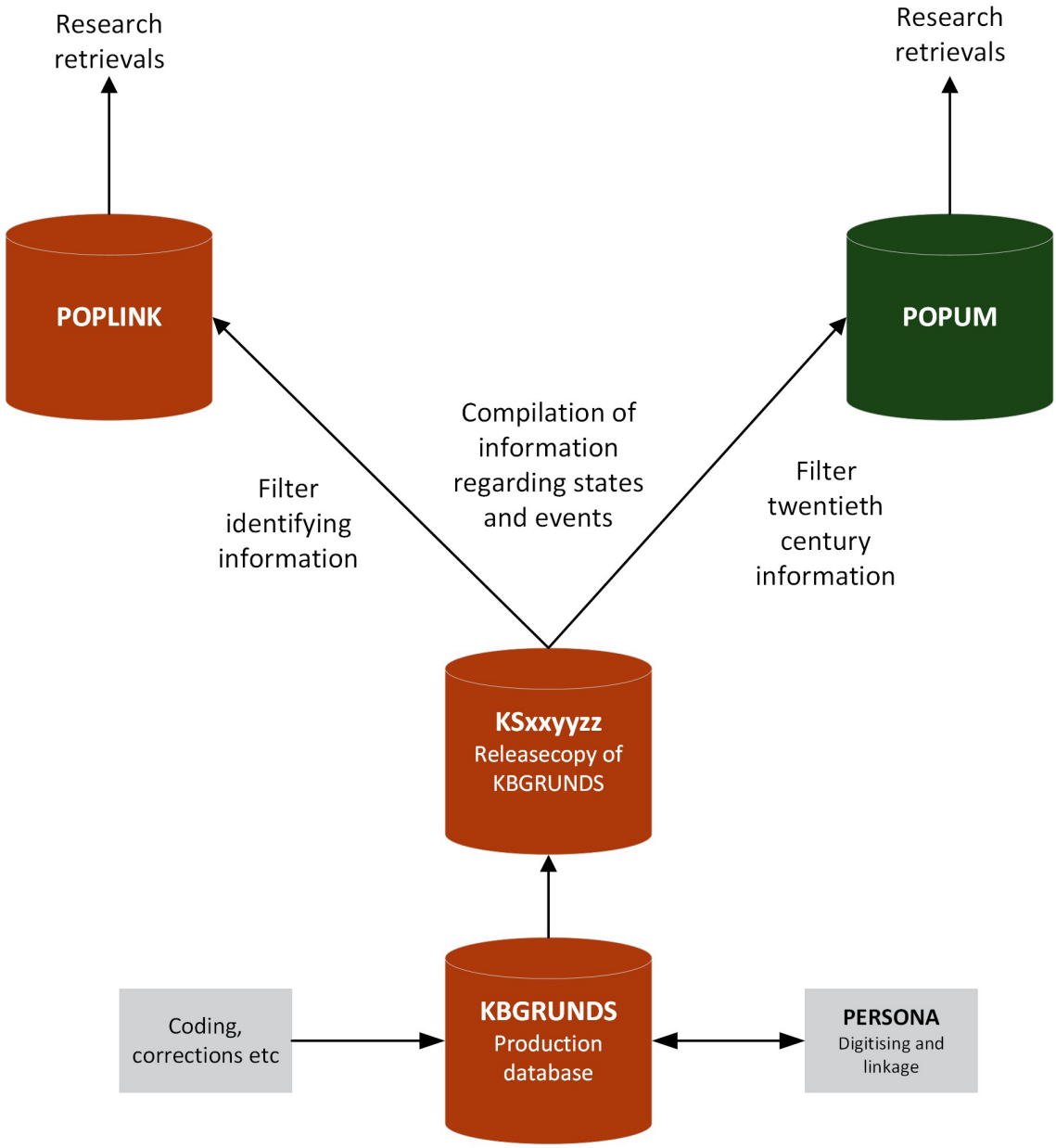
2.4 THE POPLINK DATABASE — INCREASED DEMANDS ON INFORMATION SECURITY

In 2008 the scope of the data collection at the DDB was extended. Before that, focus had been on data from the 18th and 19th centuries and data collection had usually ended around the years 1900–1910. In a new project, data from the 20th century from two regions in northern Sweden was digitized, allowing researchers to take advantage of the long time spans in the Swedish sources. The objective was to create a new asset for micro-level analysis of the processes that transformed society, through the demographic transition and beyond, by bridging the present gap between historical population databases and modern registers (Westberg, Engberg, & Edvinsson, 2016).

Handling individual-level data stretching as far as 1960 raises several ethical issues, adding a new level of demands on the information- and IT-security. Already before the implementation of the European General Data Protection Regulation[1] (GDPR) in 2018, Sweden had a strict legal framework concerning the protection of privacy, with implications for the data production process as well as for the release of data for research. The extended scope of data collection, including information that constitutes personal data according to the GDPR, meant that some changes had to be made regarding the storage of the databases. The production process including the production database KBGRUND (from now on KBGRUNDS) was moved to a secure encrypted network. A new population database, POPLINK, with the same structure as POPUM, was created in the secure network to handle retrievals containing 20th century data. The secure network safeguards data with a double layer of physical security and access authentication. The population database POPUM, containing no information about living persons, and thus no personal data, is still stored in the "open" network to make this information easier to retrieve (Westberg, Engberg, & Edvinsson, 2016). The information in the production database KBGRUNDS is filtered in different ways to the population databases POPUM and POPLINK, as shown in Figure 3. The red databases are in the secure encrypted network, and the green database in the "open" network.

1 https://gdpr.eu/

Figure 3 *Compilation of POPUM and POPLINK*



The first version of POPLINK was built by adding new data for the period 1900–c1960 to parishes already existing in the POPUM database at the DDB to reduce the time between data entry and release of the data. The linkage between old and new data worked almost seamlessly and POPLINK has since then been a valuable resource at the DDB and for the research community. Data is continuously added to the database as new parishes are digitized and linked. New versions of POPUM and POPLINK are released about once a year and as part of this process new decisions are made regarding which information must be stored in the secure encrypted environment.

2.5 THE NEW DIGITIZATION SYSTEM PERSONA AND VERSION 6 OF KBGRUNDS

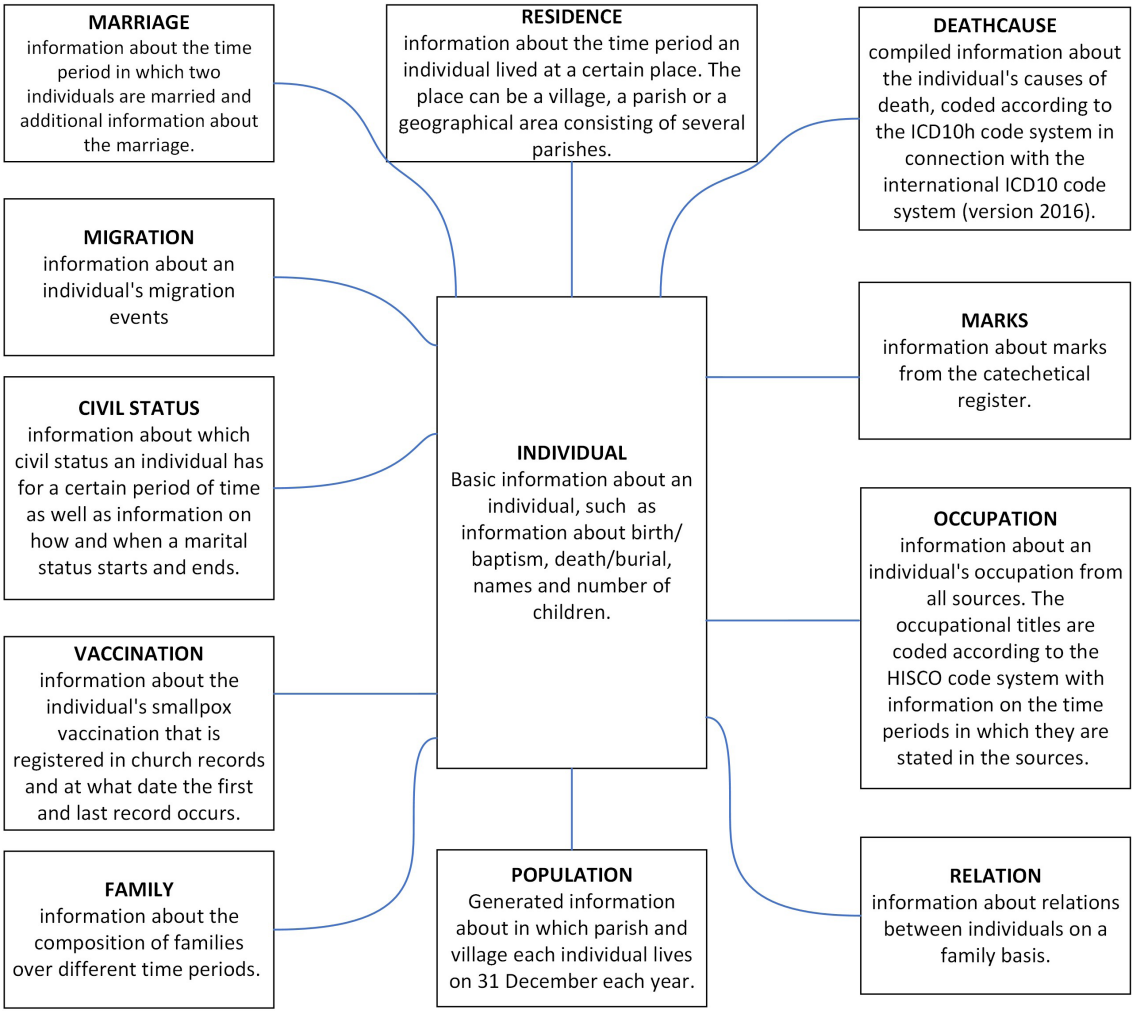
In 2012 the DDB was granted funding from the Swedish Research Council, to develop a new digitization system. The old system Regina had been used for 12 years, and during that period a number of revisions had been made to the software. Moreover, the software and programming languages used for developing REGINA were at the time becoming outdated, so it was due time to develop a new system instead of making new revisions.

But this time the aim was higher than just serving the needs of the DDB. The idea was to develop a web-based tool for database building that could serve the entire research community, not only the

DDB, increasing the interoperability of population databases in Sweden. During the project period, the software was named PERSONA. However, building a tool intended for a more general use called for a higher level of harmonization and standardization, in database modelling as well as in coding of variables, using national and international standards. Version 6 of KBGRUNDS was designed to be a strictly normalized database adapted to the new digitization system PERSONA.

The first version of PERSONA was put into operation in 2016 and has since then been used for all digitization and linkage of parish registers at DDB. The system is at present also used by the Swedish National Archives and by Gothenburg University (the Gothenburg Population Panel). All the software needed to produce a new version of POPUM and POPLINK, version 6, had to be redesigned to fit version 6 of KBGRUNDS. A major change was also made in the structure and contents of the new POPUM and POPLINK, as shown in Figure 4. As the production database is strictly normalized it is difficult to retrieve and analyze data. To facilitate the work with data retrievals, make it easier for users to retrieve data, the new versions of POPUM and POPLINK only include tables with compiled information describing certain events and states of an individual. It took some time to make these changes, but the result was worth the wait. Most research retrievals can now be made much faster than with earlier versions of the databases.

Figure 4 *POPUM and POPLINK version 6.4*



3 DEVELOPMENT OF THE LINKAGE PROCESS

From the infancy of the DDB it was evident that linkage of the digitized records would be required to create an efficient research database for life-course studies. Numerous records from different sources had to be ordered into a continuous chronology, constructing biographies, which in the ideal case cover the entire life span of an individual from the cradle to the grave. The presence of kinship links and family relationships has also significantly increased the scientific value of and usefulness of the data.

3.1 TOWARDS A SEMI-AUTOMATED PROCESS

As already mentioned, the very first form of record linkage was to order and group the paper cards with transcribed entries into bundles, one bundle representing one individual (Edvinsson & Engberg, 2020). Linkage of family relationships was also done manually, by way of a form of traditional family reconstitution. Along with the advances of computer technology, these manual procedures were replaced by semi-automated and computer-aided linkage systems.

In the late 1990s the first linkage software, ManLank (Manual Linkage), was developed. It was a computer-aided system that supported a manual linkage of individual records and of family relationships within a parish. At that time only two specially appointed personnel were allowed to work with linkage. They had no time pressure to finalize the linkage of a parish and used a lot of time scrolling backwards and forward in the sources to find matches.

Between 1999 and 2002, a semi-automated linkage process, using a combination of automated and manual methods of record linkage was developed and implemented. The most important improvement was the development of new software for automated linkage, CoreLink (Computerized Record Linkage), described in detail in section 3.3.1. Linkage was now performed by all data entry assistants, making linkage a fully integrated part of the data production process. In the first decade of the new millennium the linkage process consisted of three steps: a) an automated record linkage with CoreLink, b) a computer-aided record linkage with an updated version of ManLank, and c) a manual linkage of relationships.

Although an automated linkage phase, a), was introduced, the computer assisted manual linkage step, b), was maintained. The aim was twofold: 1) to perform linkage of records that could not be linked by the software, and, at the beginning, also 2) to validate the result from the automated record linkage process. In all kinds of linkage, secure links has always been a main priority for the DDB, far more important than a high linkage rate. For the manual linkage, all available information in the sources could be used to validate and link information that could not be linked by the software. For instance, if information about an individual was scarce, records with information about parents and partners could be tracked and used to make a secure linkage. At the beginning, all links made by CoreLink were manually scrutinized, but after some years the manual process was changed into a residual linkage, only focusing on linked records that were flagged by the system as incomplete or inconsistent and needing manual attention. The main reason for changing the process was that scrutinizing every link was very time-consuming, and, as very few changes were made, it had almost no impact on the result. Examples of links still being scrutinized, as they are flagged, are for instance individual records where the automated system has detected gaps in life-biographies, which are not associated with migration. Another common scenario is when a date for a birth, marriage or a death is noted in the longitudinal parish register, without a matching record in the event registers linked to the biography. The last step in the linkage process, c), was a linkage of relations, that is, linking parents to children and spouses to each other. This linkage step was performed manually with computer assistance. All relation links were stored in a specific relation table in the database, a table which still is instrumental for constructing genealogies over several generations.

3.2 FURTHER LINKAGE STEPS AND INCREASED AUTOMATIZATION

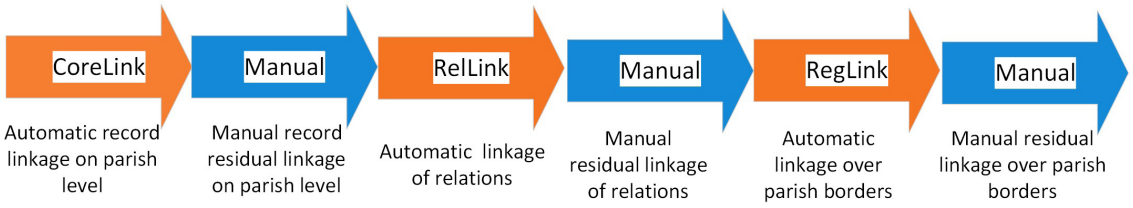
Since the digitization and linkage of parish records mostly have been performed register by register for one parish at a time, a high linkage rate on the parish level does not necessarily imply complete life-course data at the individual level. It is particularly migration back and forth over parish borders that causes problems. The registers do not cover events that took place in other parishes before or after migration, and therefore most migrants have incomplete life course data or apparent gaps in their life-biographies. Another problem is caused by administrative changes in parish structure, resulting

in that a distinct part of the population disappears from the sources for other reasons than migration at a fixed point in time. A problem that became more and more evident as the database expanded was duplets, caused by short-distance migration, that is, people moving between adjacent parishes, sometimes several times. Some of these migrants ended up having two or more different identities in the database: one in their parish of origin and other ones in parishes they later belonged to. A number of steps have been taken to solve these issues.

The problem with information loss due to short distance migration and administrative changes of parish borders was solved by introducing an additional linkage step, linking larger areas consisting of adjacent parishes. Accounting for short distance migration significantly increases the number of complete life-courses in the data. At first, this linkage step was performed manually using SirLink (Simple Regional Linkage), an in-house developed software. However, it ended up being a time-consuming enterprise and soon it became obvious that assistance from an automated linkage software was needed. Since CoreLink, was focused on linking records within a parish and this kind of linkage needed a completely different approach, it was decided to develop new software. In 2010, new software called RegLink (Region Linkage) was added to the linkage process allowing for automatic linkage of individuals across parishes.

In 2014, the most recent addition to the linkage process was made when RelLink (Relation Linkage) an automatic software for linking relations, was developed. The linkage process, as shown in Figure 5, was thereby completed with both automated and manual steps for all three types of linkage. With these new automatic steps, the linkage process time was significantly reduced. In 2016 finally, all manual steps of linkage were included in the same software, the production system PERSONA, described above.

Figure 5 *Current Linkage Process*



3.3 SOFTWARE FOR AUTOMATED LINKAGE – TECHNICAL ASPECTS

The development of automated linkage software at DDB described in 3.1-3.2 began with an evaluation of different linkage approaches. A probabilistic approach, as used when linking American censuses (see for example Abramitzky, Boustan, Eriksson, Feigenbaum, & Pérez, 2021), was compared to a rule-based approach. Different probabilistic methods were also tested.

The data from Swedish church records is of high quality and has, as described in section 2.1, the advantage of the records being "pre-linked" by the minister who kept the books. The longitudinal parish registers have references to information in the event registers, and the event registers include distinct references to the longitudinal parish registers. The latter also include individual level references to previous and subsequent registers, making it possible to trace individuals between the registers, back and forth in time (Nilsdotter Jeub, 1993, p. 65). Therefore, it was decided that a rule-based approach where the references in the sources between registers could be used was to be preferred. The rule-based approach resulted in a high linkage rate with a very small amount faulty links.

3.3.1 CORELINK — AUTOMATED RECORD LINKAGE WITHIN A PARISH

In 2002 the first version of the software CoreLink for automated linkage was finalized and implemented in the production process. In this first version, four out of the five principal sources in the Swedish parish records were linked together, namely the longitudinal parish records, birth, marriage, and death registers. The migration registers were left out as they were considered too difficult to handle automatically, due to their frequently poor information about family members. Often, the registers included only the name of the head of the migrating family, or group of migrants, along with information about the number of men and women that were moving out or in together. The solution became to manually

add a record containing the same information to all migrants in the family group. In the second version of Core Link implemented in 2009, migration records were included in the automated linkage process. This version linked only the head of the migration group not the other individuals, because only the head of the group had enough information to make automated linkage possible. In the third version of the software, released in 2012, the unique national registration number, introduced in 1947–48 was included as a linkage variable, something that significantly improved the automated linkage of 20th century records.

As stated before, the linkage is rule based, and the primary aim of CoreLink is to use well defined algorithms and rules to decide if a link can be established or not. Matching is done between pairs of records or between groups of records. Exact matches and clear miss-matches are processed automatically using well-tried algorithms for searching and matching. The software gives all individuals in the data unique identity numbers. If there are records impossible to link to any other record, a new individual will be "created" with only one record.

The core of the Swedish parish registers are the longitudinal registers, with their comprehensive system of references between individual records and sources. They are therefore used as the main source for linkage. During the development of CoreLink, different sequences of linking event registers to the longitudinal registers were tested, to determine which sequence resulted in the best links. This resulted in that the following sequence was implemented: a) linking birth and baptism records to the longitudinal parish records; b) matching and linking together all individual records in the longitudinal registers; c) matching and linking migration records; d) matching and linking banns and marriage records, and finally, e) matching and linking death and burial records.

Since a person in a population often cannot uniquely be identified by his or her name — naming practices showed very little variation before 1900 — it was necessary to include other unique variables in the linkage. In the chosen model, sex was used as a blocking variable together with either year of birth or year of an event, such as year of marriage. The other variables used in different combinations were date of birth, parish of birth, date of event (death, marriage, migration) and page references made by the minister. Names are compared, using Levenshtein distance (https://www.cuelogic.com/blog/the-levenshtein-algorithm) and standardized names.[2] Since most common first names in Sweden are short, working with standardized names has been important to reach the 75% correspondence between the strings, that according to the definition is required for a match. A good example is the name PER, which also can be spelled PÄR or PEHR. Since in both cases the difference between the name variants is one letter out of three, the correspondence only reaches 66%, a measure that is considered too low to constitute a match. This is the reason why name standardization considerably improves the linkage rate

Documenting how a link has been made is important and therefore every decision made by the software was entered into a log table stating which rule a certain record was linked by. After the automated linkage, a special routine is executed to identify problems in the links such as illogical start or end dates, overlapping records, illogical gaps of time between records, the occurrence of more than one birth parish, errors in references between sources, etcetera. This information is used during the manual validation of the linkage, identifying possible faulty links.

3.3.2 RELLINK — AUTOMATED LINKAGE OF RELATIONS

The Swedish parish registers offer ample information about relationships. In the longitudinal parish registers families were registered together, and the type of relation between the family members is normally specified for all individuals on each page. Parents are named in the registers of birth and death. Names of spouses and sometimes also parents are usually found in both marriage and death registers. As most individuals have multiple records, there is a need to create unique relation links. In 2014 the software RelLink for automated linkage of relations between parents and children and between spouses was developed and implemented. The software uses several different rules for each kind of relation. For example, information from birth records is used to determine whether a relation between a child and a parent is of biological nature. Information from marriage records and longitudinal parish records is used to define status for the relation between spouses: engaged, betrothed or married.

2 The Levenshtein distance is a metric for measuring the difference between two sequences.

3.3.3 REGLINK — AUTOMATED LINKAGE ACROSS PARISHES

The RegLink software matches and links individuals in different parishes, into one unique identity, making it possible to follow individuals moving from one parish to another within the same area in the data.

RegLink was constructed in a similar fashion as CoreLink, but includes a new, separate group of rules for linking individuals in different parishes. The rule-based linkage in RegLink is based on a number of identified scenarios. The first scenario uses the events, birth, marriage and death, from three different aspects.

1. If an individual has a record in a source in parish A, indicating that a birth, a marriage or a death took place in parish B, the software tries to find this individual in parish B, as noted in the register.
2. If an individual has information about birth, marriage or death in the longitudinal parish register where he/she is registered, without a matching record in the birth-, marriage-, or death register in the same parish, the software tries to find these records in other parishes.
3. If an individual in a birth, marriage or death record in parish A, is mentioned to reside in parish B, the software tries to locate this individual in a register from parish B.

The second scenario uses information about migration from migration registers and longitudinal parish registers. Links are made when an individual is leaving one parish and entering another parish in the same year, and with the same co-movers. When using this scenario surname is an important variable. Women, who frequently migrated in association with marriage and thus also might have changed their surnames, must be handled in a special way. If the date of marriage is close to the time of the migration, both the womans maiden name and her new husband's surname are used as linking variables.

The third scenario covers administrative migration, that is, when an individual is transferred to another parish due to an administrative change, such as the detachments of a new parish or changed parish borders. Here, linkage requires that the individual is recorded at the same place of residence and with the same relatives in the parish register, both before and after the administrative migration, and that the year of the administrative migration is equivalent to the start year of the new parish.

A particular challenge has been handling individuals with periods of absence. One of the most difficult groups to link are single men and women moving out of the region covered by the database and later returning to a different parish than the one they once left. Trying to solve this problem, a set of rules looking for the absence of records of residence in the life-biography of an individual were added, aiming to find matching individuals for at least a part of that absence. Other variables used in these cases are date of birth, names, parish of birth, occupation and relatives. The linkage of individuals in this group has been improved by the automated linkage, but a small under-linkage remains.

4 CONCLUSIONS

During the 50 years that DDB has existed, the purpose of the infrastructure has remained the same: building high-quality longitudinal population databases for research, developing effective methods for database construction, and disseminating data for research. From the 1970s and until today, technological and methodological advancements and innovations within this field have been immense. Manual excerpts and rudimentary forms of linkage have been replaced by comprehensive digitization systems with processes, at the same time reducing the process time and improving the quality of the data. Early database models with limitations have given way for advanced and flexible database management systems, increasing data security and consistency and facilitating long-term management and data retrievals. A certain amount of manual validation of the results of the automated processes has however been kept ensuring the quality of data. Even though the digitization systems have improved, the actual interpretation of the text is still manual. Looking forward, handwritten text recognition (HTR) techniques stand out as an interesting development of the technical infrastructure, following the example of the BALSAC in Canada (Vézina & Bournival, 2020). Another important improvement is the national partnership within the SwedPop infrastructure, that besides its large value for future comparative research also has brought about valuable methodological collaborations with other Swedish databases, for example SEDD at Lund University.

REFERENCES

Abramitzky, R., Boustan, L., Eriksson, K., Feigenbaum, J., & Pérez, S. (2021). Automated linking of historical data. *Journal of Economic Literature,* 59(3), 865–918. doi: 10.1257/jel.20201599

Brändström, A. (1984). *"De kärlekslösa mödrarna": Nedgången i spädbarnsdödlighet i Sverige under 1800-talet, med särskild hänsyn till Nedertorneå* (Doctoral dissertation). Umeå University.

Brändström, A. (2009). Demografiska databasen och historisk demografi i Umeå — "En allvarlig felinvestering"? In R. Jacobsson (Ed.), *Thule: Kungliga Skytteanska Samfundets årsbok 2009* (pp. 211–222). Umeå: Kungliga Skytteanska Samfundet.

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM, 13*(6), 377–387. doi: 10.1145/362384.362685

Edvinsson, S. (2000). The Demographic Data Base at Umeå University: A resource for historical studies. In P. Kelly Hall, R. McCaa, R., & G. Thorvaldsen (Eds.), *Handbook of international historical microdata for population research (*pp.231–248). Minneapolis: Minnesota Population Center. Retrieved from https://international.ipums.org/international/microdata_handbook.shtml

Edvinsson, S., & Engberg, E. (2020). A database for the future. Major contributions from 47 years of database development and research at the Demographic Data Base. *Historical Life Course Studies, 9*, 173–196. doi: 10.51964/hlcs9305

Johansson, E., & Åkerman, S. (1973). "Faktaunderlag för forskning. Planering av en demografisk databas". *Historisk tidskrift, 3*, 406–414.

Karlsson, T., & Lundh, C. (2015). *The Gothenburg Population Panel 1915–1943: GOPP version 6.0.* (Papers in Economic History No. 18). Lund University Publications. Available from https://lup.lub.lu.se/search/publication/7870561

Kesztenbaum, L. (2021). Strength in numbers. A short note on the past, present and future of large historical databases. *Historical Life Course Studies, 10*, 5–8. doi: 10.51964/hlcs9557

Lund, R. (2017). *Regler för konvertering av KBGRUNDS5 till KBGRUNDS6*. Centre for Demographic and Ageing Research, Umeå Universitet.

Mandemakers, K., & Dillon, L. (2004). Best practices with large databases on historical populations. *Historical Methods, 37*(1), 34–38. doi: 10.3200/HMTS.37.1.34-38

Nilsdotter Jeub, U. (1993). *Parish records: 19th century ecclesiastical registers.* Information from the Demographic Data Base. Umeå: Umeå University, Demographic Data Base.

Vézina, H., & Bournival, J.-S. (2020). An overview of the BALSAC population database. Past developments, current state and future prospects. *Historical Life Course Studies, 9*, 114–129. doi: 10.51964/hlcs9299

Vikström, P., Edvinsson, S, & Brändström, A. (2006). Longitudinal databases-sources for analyzing the life-course: Characteristics, difficulties and possibilities. *History and Computing, 14*(1–2), 109–128.

Westberg, A, Engberg, E., & Edvinsson, S. (2016). A unique source for innovative longitudinal research: The POPLINK database. *Historical Life Course Studies, 3*, 20–31. doi: 10.51964/hlcs9351