# Historical Population Database of Transylvania. Sources, Particularities, Challenges, and Early Findings

By Luminița Dumănescu, Mihaela Hărăguș, Angela Lumezeanu, Elena Crinela Holom, Nicoleta Hegedűs, Daniela Mârza, Diana Covaci and Ioan Bolovan

To cite this article: Dumănescu, L., Hărăguș, M., Lumezeanu, A., Holom, E. C., Hegedűs, N., Mârza, D., Covaci, D., & Bolovan, I. (2022). Historical Population Database of Transylvania. Sources, Particularities, Challenges, and Early Findings. *Historical Life Course Studies, 12*, 133–150. https://doi.org/10.51964/hlcs12038

HISTORICAL LIFE COURSE STUDIES

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

VOLUME 12, SPECIAL ISSUE 5,
2020

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.

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

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

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

HISTORICAL LIFE COURSE STUDIES
VOLUME 12 (2022), published 28-06-2022

# Historical Population Database of Transylvania

## Sources, Particularities, Challenges, and Early Findings

| | |
|---|---|
| Luminița Dumănescu | Centre for Population Studies, Babeș-Bolyai University, Cluj-Napoca |
| Mihaela Hărăgus, | Centre for Population Studies, Babeș-Bolyai University, Cluj-Napoca |
| Angela Lumezeanu | Centre for Population Studies, Babeș-Bolyai University, Cluj-Napoca |
| Elena Crinela Holom | Centre for Population Studies, Babeș-Bolyai University, Cluj-Napoca |
| Nicoleta Hegedűs | Centre for Population Studies, Babeș-Bolyai University, Cluj-Napoca |
| Daniela Mârza | Center for Transylvanian Studies, Romanian Academy, Cluj-Napoca |
| Diana Covaci | Centre for Population Studies, Babeș-Bolyai University, Cluj-Napoca |
| Ioan Bolovan | Centre for Population Studies, Babeș-Bolyai University, Cluj-Napoca |

## ABSTRACT

The Historical Population Database of Transylvania (HPDT) is a research tool for population studies developed since 2014 at the Centre for Population Studies in Cluj-Napoca, financed by an SEE-Norway Grant. HPDT employs a source-oriented approach for recording data from the parish registers kept by the Transylvanian churches, focusing primarily on the main vital events such as births, marriages, and deaths. The data entry process was followed by the standardization of various information, such as names, occupations, locations and causes of death, thus allowing the initiation of a linkage process. The database has already been employed in a wide-ranging series of analyses conducted on datasets extracted from HPDT, which include infant and adult mortality, nuptiality and age at first marriage, social mobility, and the medicalization of childbirth. The wealth of information it includes will enable many more scientific investigations.

**Keywords:** Historical database, HPDT, Transylvania, Parish registers, Historical demography

e-ISSN: 2352-6343
DOI article: https://doi.org/10.51964/hlcs12038

# 1 INTRODUCTION

Historical demography began its journey in Romania in the interwar period and grew in the shadow of other sciences. Since the 1960s, case studies and microlevel analyses based on the Henry and Fleury method of family reconstitution have enriched the Romanian historical demography. Only recently this research has developed a population database constructed according to the most recent methodological principles (Mandemakers & Dillon, 2004) and aiming for full compatibility with the latest version of the *Intermediate Data Structure* (IDS; Alter & Mandemakers, 2014). The Historical Population Database of Transylvania (HPDT) was developed by the Centre for Population Studies in Cluj-Napoca. Data entry started in 2014, after the project received financial support and was developed together with the Norwegian Historical Data Centre (Tromsø, Norway). A team of almost 20 people — researchers and data entry operators — started to collect, transcribe, standardize and link data from a sample of around 7% of the historical population of Transylvania, and they delivered a working database compatible with the existing international standards. After the project's completion in April 2017, researchers of the Centre for Population Studies continued to clean, standardize, link and add data to HPDT. The purpose of this paper is to present the Transylvanian database, one of the newest databases in Europe, focusing on its particularities and on the challenges encountered during its development and implementation.

Development of the database was based on the source-oriented approach and attempted to faithfully replicate all types of data found in the parish registers, including the smallest possible details. The database was constantly adapted to integrate new fields as soon as new types of data emerged, thus reflecting the complexity and the diversity of the sources.

The Historical Population Database of Transylvania (HPDT) provides a tool for researching demographic phenomena at the micro level in Transylvania, the northern part of present-day Romania (see Figure 5). The intimate details revealed by the newly available data allowed scholars to see a world that functioned differently than previously thought, when only macro-level data were available (census, aggregates). These sources shaped a social history which sheds new light on the old world — often considered obsolete and traditional — of a province situated at the periphery of the Austrian-Hungarian Empire (until it became part of Romania in 1919).

The main sources of the HPDT are the parish registers kept by churches and covering the period between 1850 and 1914–1920. These sources recorded the most important events in an individual life course from a demographic perspective. The architecture of the database was conceptualized by starting with the three principal registers: those for the baptism, marriage, and burial events. Betrothals were added to the database where they were available.

The structure of the database is based on the main events from the parish registers: births, marriages, burials. Each main table incorporates all the fields that can be found in the register for the respective event regardless of the denomination of the register. Each data entry form reflects the structure of the corresponding table from the database. Because of the complexity of the sources the architecture of the database steadily evolved during the project implementation.

The Historical Population Database of Transylvania is a relational database, implemented in MySql, an open source database management system. There are two main components of HPDT: a research database on the one hand, and a public open access one on the other. The research component is built on three elements: a source database, a standard database and a third database which contains linked individuals. The public database (http://hpdt.ro:4080) is an open access website that offers insights on the Transylvanian population to a broader public. The structure of the database has been extensively described in the doctoral thesis of Angela-Cristina Lumezeanu (2019).

The following sections address the particularities of parish registers in Transylvania, sampling strategy and sample composition, principles of data entry, and ongoing development of the database structure in response to the heterogeneity of the information found in the parish registers. We follow the discussion of the database with an overview of the main results of research based on HPDT data.

# 2 SOURCES AND DATA ENTRY

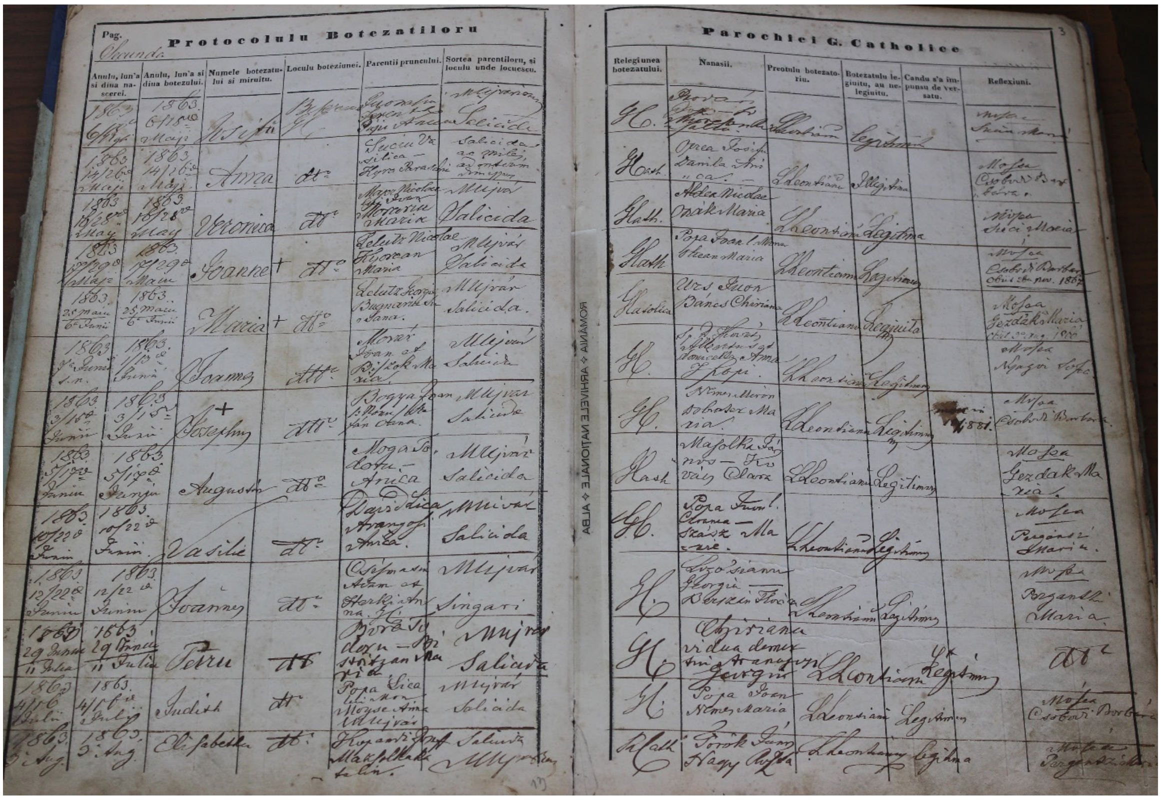
## 2.1 SOURCES

The main sources of the HPDT are the parish registers that recorded life events such as birth of a child (in Baptism registers), wedding (in Marriages registers), engagements (in Betrothal registers) and deaths (in Burial registers). Until 1895, when Hungarian Law enforced the civil registration of the life events, these parish registers could be regarded as official records, and, despite all their limits and fragmentary character, they appear to be reliable sources for historical demography. The HPDT is also based on parish registers after 1895, since inclusion of civil certificates would have required a new structure of the database.

Church registers in Transylvania were written in several languages (Romanian, Hungarian, Latin, German, etc.), with different alphabets (Latin, Cyrillic, etc.), and came from diverse denominations: Orthodox, Greek-Catholic, Roman-Catholic, Reformed, Lutheran and Jewish. This diversity translated into different spellings of names and variation in the registers' headings, which required multiple adaptations of the database structure (discussed in next sections). The following Figures 1–4 give some examples of these sources, illustrating the structural heterogeneity of the parish registers.

Information about births was derived from the baptismal registers (see Figures 1 and 2). The information was organized in three sections with fields for data regarding the child, his parents and the baptism respectively. Some of these are text fields, for the name of the child and of the parents, the birth place, the baptism place, the parents' occupations and residence. In some parish registers the data was minimal with only the more usual categories of information including the date of the baptism, the name of the child, the parents' names (in some cases only the father is mentioned), the godparents, and the name of the priest who performed the baptism. Other sources held a wealth of additional information about the occupation of the parents, their age, their other relatives, the address of residence, the midwife, the age, status and occupation of the godparents, about vaccination (the date thereof, the name of the physician performing the inoculation) and even about the death of the baptized (sometimes filled in decades after the baptism record). In addition to the general categories of information, there were specific ones such as: legitimacy of the child, whether the child was the result of a single or multiple birth, if he or she was anointed or not (one of the sacraments).

Figure 1 *Excerpt from a baptism register, 1863 (Latin)*



Pag. Secunda

Protocolulu Botezatiloru

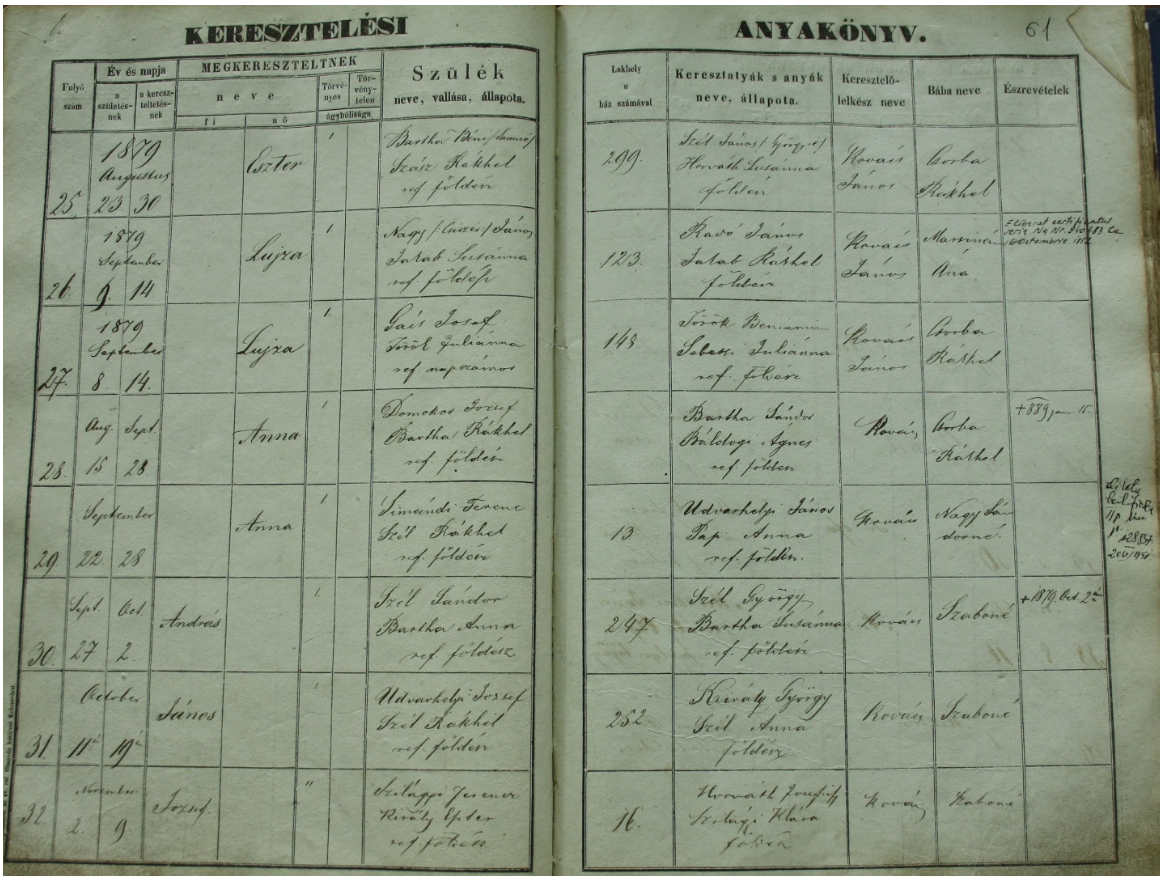
Parochiei G. Catholice

3

| Anulu, lun'a si dina nascerei. | Anulu, lun'a si dina botezului. | Numele botezatului si miruitu. | Loculu boteziunei. | Parentii pruncului. | Sortea parentiloru, si loculu unde locuescu. | Relegiunea botezatului. | Nanasii. | Preotulu botezatoriu. | Botezatulu legiuitu, au nelegiuitu. | Candu s'a impunsu de versatu. | Reflexiuni. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1863 | 1863 | Iosifu | | | | | | | | | |
| 1863 | 1863 | Anna | dto | | | | | | | | |
| 1863 | 1863 | Veronica | dto | | | | | | | | |
| 1863 | 1863 | Ioanne | dto | | | | | | | | |
| 1863 | 1863 | Maria | dto | | | | | | | | |
| 1863 | 1863 | Ioanne | dto | | | | | | | | |
| 1863 | 1863 | Iosifu | dto | | | | | | | | |
| 1863 | 1863 | Augustin | dto | | | | | | | | |
| 1863 | 1863 | Vasile | dto | | | | | | | | |
| 1863 | 1863 | Ioanne | dto | | | | | | | | |
| 1863 | 1863 | Petru | dto | | | | | | | | |
| 1863 | 1863 | Iudith | dto | | | | | | | | |
| 1863 | 1863 | Elisabetha | dto | | | | | | | | |

*Note: The columns contain the following information: the year, month, day for the birth and baptism, the first name of the baptised child, place of baptism, the parents of the child, their occupation and place of living, child's denomination, the name of the godparents, the name of the priest, vaccination date and reflections.*

Figure 2 *Excerpt from a baptism register, 1879 (Hungarian)*



6 KERESZTELÉSI ANYAKÖNYV. 61

| Folyó szám | Év és napja: a születésnek | Év és napja: a kereszteltetésnek | Megkereszteltnek neve: fi | Megkereszteltnek neve: nő | Törvényes ágybólisága | Törvénytelen ágybólisága | Szülék neve, vallása, állapota | Lakhely a ház számával | Keresztatyák s anyák neve, állapota | Keresztelő-lelkész neve | Bába neve | Észrevételek |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 25. | 1879 Augustus 23 | 30 | | Eszter | 1 | | Bartha Benjámin<br>Szász Rákhel<br>ref. földm. | 299. | Szél János (György)<br>Horváth Susánna<br>földm. | Kovács János | Gorba Rákhel | |
| 26. | 1879 September 9. | 14 | | Lujza | 1. | | Nagy (Császi) János<br>Salab Susánna<br>ref. földm. | 123. | Radó János<br>Salab Rákhel<br>földm. | Kovács János | Marina Ána | Eliberat certificatul seria Na Nr. 940483 la 16 octombrie 1952 |
| 27. | 1879 September 8. | 14. | | Lujza | 1. | | Gál József<br>Török Julianna<br>ref. napszámos | 148 | Török Benjámin<br>Sebess Julianna<br>ref. földm. | Kovács János | Gorba Rákhel | |
| 28. | Aug. 15 | Sept. 28 | | Anna | 1 | | Domokos József<br>Bartha Rákhel<br>ref. földm. | | Bartha Sándor<br>Báldogi Ágnes<br>ref. földm. | Kovács | Gorba Rákhel | +889 jan. 15. |
| 29. | September 22. | 28. | | Anna | 1 | | Simándi Ferenc<br>Szél Rákhel<br>ref. földm. | 13. | Udvarhelyi János<br>Pap Anna<br>ref. földm. | Kovács | Nagy Sándorné | [illegible] 12887 200/951 |
| 30. | Sept. 27 | Oct. 2. | András | | 1. | | Szél Sándor<br>Bartha Anna<br>ref. földm. | 247 | Szél György<br>Bartha Susánna<br>ref. földm. | Kovács | Szabóné | +1879. Oct. 2. |
| 31. | October 11. | 19. | János | | 1 | | Udvarhelyi József<br>Szél Rákhel<br>ref. földm. | 252. | Király György<br>Szél Anna<br>földm. | Kovács | Szabóné | |
| 32. | November 2. | 9 | József | | " | | Szilágyi Ferenc<br>Király Eszter<br>ref. földm. | 16. | Horváth József<br>Szilágyi Klára<br>földm. | Kovács | Szabóné | |

*Note: The columns contain the following information: the year, month, day for the birth and baptism, the child's name (different table for boys and girls), legitimacy, the name of the parents and their social statute, the house number, name, surname and statute for godparents, the name of the priest, the name of the midwife, observations.*

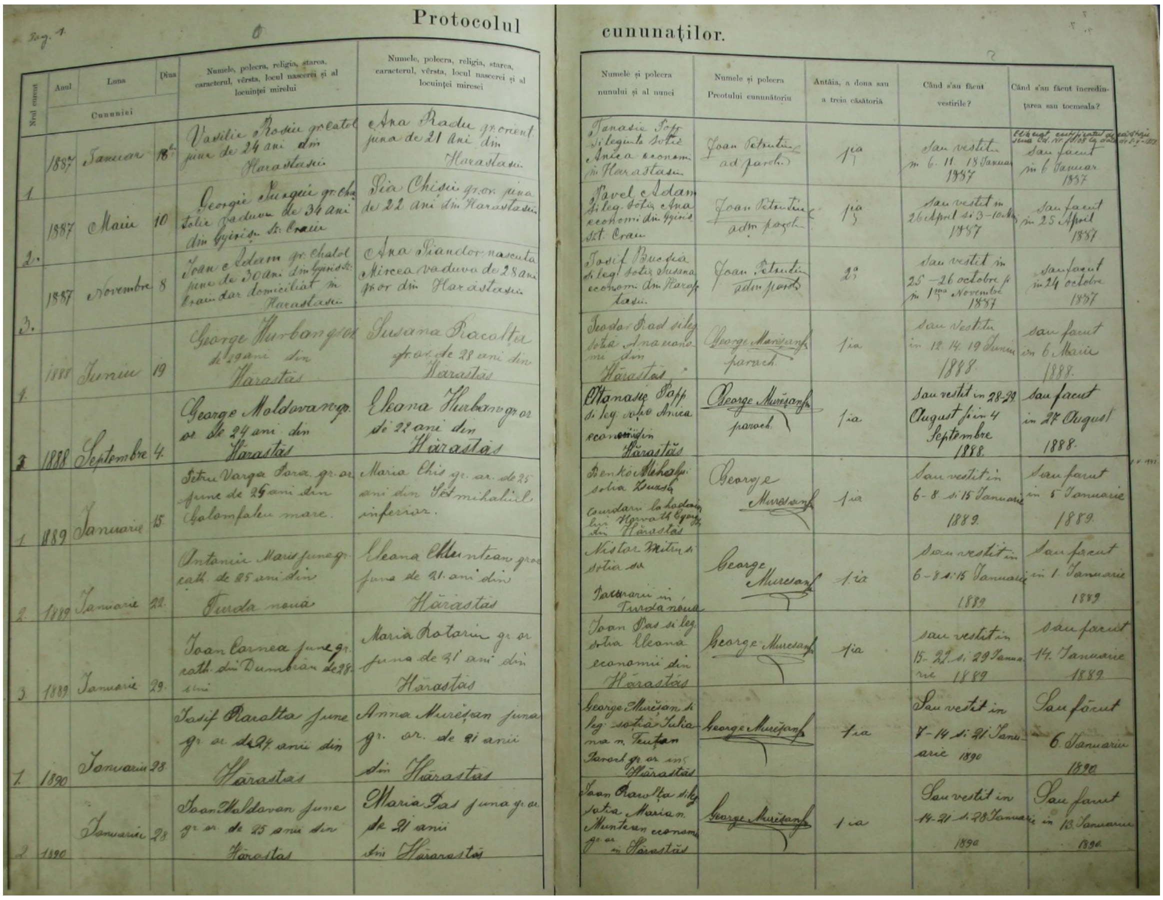
Marriages registers (see Figure 3) accounted numerous actors involved in the event, such as bride and groom, their parents, and godparents (who could be multiple pairs). All need fields for names, civil status, denomination, age, occupation etc. Therefore, the information from this type of source resulted in the largest number of fields in the database.

The Deaths registers (see Figure 4) contained information such as the name of the deceased, denomination, marital status, occupation, residence, birth date, death and burial dates, cause of death, information about parents or spouse, as well as the priest who registered the event. Some registers included more information, such as the legitimacy status if the deceased was a child, the date of the death certificate or mentioned different relatives of the deceased.

Information from the Betrothals registers, which are particular to Orthodox and Greek Catholic Churches, was also included. Betrothals describe the future bride and groom, their parents or guardians (names, literacy, denomination). Although they are interesting life course events, betrothals were usually not recorded, so betrothal registers are not available for every marriage register included in the HPDT. Thus, it was decided to include betrothals in the database, but not to use them for standardization and linkage.

Data entry started by filling a datasheet with information about the source itself. The fields describe county, parish, denomination, type of event, language, alphabet, the dates of first and last records. This datasheet also has a field for the stage of processing of the source (transcribed, checked, standardized), which is updated at later stages.

Figure 3 *Excerpt from a marriage register, 1887 (Latin)*



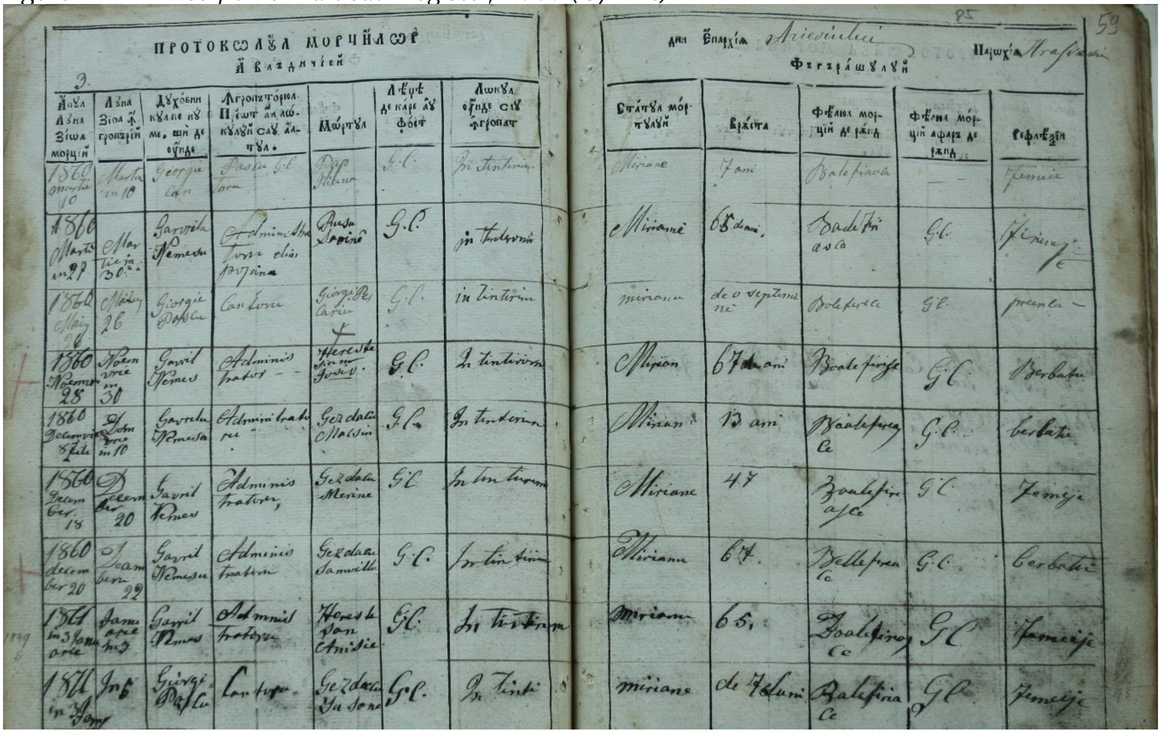
Pag. 1.

Protocolul cununaţilor.

| Nrul curent | Anul | Luna | Dina | Numele, polecra, religia, starea, caracterul, vêrsta, locul nascerei şi al locuinţei mirelui | Numele, polecra, religia, starea, caracterul, vêrsta, locul nascerei şi al locuinţei miresei | Numele şi polecra nunului şi al nunei | Numele şi polecra Preotului cununătoriu | Antâia, a doua sau a treia căsătoriă | Când s'au făcut vestirile? | Când s'au făcut încredinţarea sau tocmeala? |
|---|---|---|---|---|---|---|---|---|---|---|
| | Cununiei | | | | | | | | | |
| 1. | 1887 | Januar | 18 | Vasilie Rodiu gr. catol june de 24 ani din Harastasu | Ana Radu gr. orient. juna de 21 Ani din Harastasu | Tanasie Popp si leguita sotie Anica economi in Harastasu | Ioan Petrutiu ad paroh | 1ia | S'au vestit in 6. 11. 18 Januar 1887 | Eliberat certificatul de căsătorie sub Nr. 1108 la data de [illegible]<br>S'au facut in 6 Januar 1887 |
| 2. | 1887 | Maiu | 10 | Georgie Jurgiu gr. chatolic vaduvu de 34 ani din Gyirisu St. Craiu | Sia Chişiu gr. or. juna de 22 ani din Harastasu | Pavel Adam si leg. sotia Ana economi din Gyirisu St. Craiu | Ioan Petrutiu adm parohu | 1ia | S'au vestit in 26 April si 3-10 Maiu 1887 | S'au facut in 25 April 1887 |
| 3. | 1887 | Novembre | 8 | Ioan Adam gr. chatol june de 30 ani din Gyiris St. Craiu dar domiciliat in Harastasu | Ana Siandor nascuta Mircea vaduva de 28 ani gr. or din Harastasu | Iosif Bucsia si leg. sotie Susana economi din Harastasu | Ioan Petrutiu adm parohu | 2a | S'au vestit in 25 - 26 octobre si in 1ma Novembre 1887 | S'au facut in 24 octobre 1887 |
| 4. | 1888 | Juniu | 19 | George Hurban gr. or. de 29 ani din Hărastăs | Susana Racalţa gr. or. de 28 ani din Hărastăs | Teodor Rad si leg. sotia Ana economi din Hărastăs | George Mureşan parach | 1ia | S'au vestit in 12. 14. 19 Juniu 1888 | S'au facut in 6 Maiu 1888. |
| 5. | 1888 | Septembre | 4. | George Moldovan gr. or. de 24 ani din Hărastăs | Eleana Hurban gr. or de 22 ani din Hărastăs | Atanasie Papp si leg. soţie Anica economi din Hărastăs | George Mureşan paroch | 1ia | S'au vestit in 28-29 August si in 4 Septembre 1888. | S'au facut in 27 August 1888. |
| 1. | 1889 | Januarie | 15. | Petru Varga Sara gr. or june de 26 ani din Galomfaleu mare. | Maria Chis gr. or. de 25 ani din Sânmihaiul inferior. | Benkö Mihaly si solia Zsuzsa courdaru la hodaiu lui Horvath Gyorgy din Hărastăs | George Mureşan | 1ia | S'au vestit in 6-8 si 15 Januarie 1889. | S'au facut in 5 Januarie 1889. |
| 2. | 1889 | Januarie | 22. | Antoniu Mariş june gr. cath. de 25 ani din Turda nouă | Eleana Muntean gr. or juna de 21. ani din Hărastăs | Nistor Mitru si sotia sa Pacurariu in Turda nouă | George Mureşan | 1ia | S'au vestit in 6-8 si 15 Januarie 1889 | S'au facut in 1. Januarie 1889 |
| 3. | 1889 | Januarie | 29. | Ioan Carnea june gr. cath. din Dumbrău de 28 ani | Maria Rotariu gr. or juna de 21 ani din Hărastăs | Ioan Pas si leg. sotia Eleana economii din Hărastăs | George Mureşan | 1ia | s'au vestit in 15- 22. si 29 Januarie 1889 | s'au facut 14. Januarie 1889 |
| 1. | 1890 | Januariu | 28. | Iosif Racolta june gr. or. de 24. anii din Hărastăs | Anna Mureşan juna gr. or. de 21 anii din Hărastăs | George Mureşan si leg. soţia Iuliana n. Teuţan Paroch gr. or ins Hărastăs | George Mureşan | 1ia | S'au vestit in 7-14 si 21 Januarie 1890 | S'au făcut 6. Januariu 1890 |
| 2. | 1890 | Januariu | 28. | Ioan Moldovan june gr. or. de 25 anii din Hărastăs | Maria Pas juna gr. or. de 21 anii din Hăranastăs | Ioan Racolţa si leg. soţia Maria n. Muntean economi gr. or. in Hărastăs | George Mureşan | 1ia | S'au vestit in 14-21 si 28 Januarie 1890 | S'au facut in 13. Januariu 1890 |

Note: The columns contain the following information: name of the deceased, denomination, marital status, occupation, residence, birth date, death and burial dates, cause of death, information about parents and spouse.

Figure 4 *Excerpt from a death register, 1860 (Cyrillic)*



ПРОТОКОЛꙊЛ МОРЦИЛОР [illegible]

дин Епархїа Aiesiului Фъгърашꙋлꙋй Парохїа Harastasu

| Анꙋл лꙋна зїоа морцїй | Лꙋна зїоа ꙟгропърїй | Дꙋховникꙋл нꙋме, ши де ꙋнде | Ꙟгропъторюл. Преѡт ал ачелꙋй сaꙋ ал алтꙋл | Мортꙋл | Легеа де каре аꙋ фост | Локꙋл ꙋнде саꙋ ꙟгропат | Статꙋл мортꙋлꙋй | Връста | Фелюл морцїй де рѫнд | Фелюл морцїй афаръ де рѫнд | Рефлексїи |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1860 martie 10 | Martie in 10 | Georgie Can | Pasku G.C. Toma | Pop Ileana | G.C. | In tintirim | Miriane | 7 ani | Boale fireasca | — | femeie |
| 1860 Martie in 27 | Martie in 30 | Gavrila Nemesu | Administratore elias Pojina | Rusu Savina | G.C. | in tintirim | Miriane | 65 de ani | Boale fireasca | G.C. | femeie |
| 1860 Maiu 26 | Maiu 26 | Giorgie Poplu | Cantore | Giorgie Pescariu | G.C. | in tintirim | miriana | de o septemâne | Boale fireasca | G.C. | prunc |
| 1860 Noemvrie 28 | Noemvrie in 30 | Gavril Nemes | Administrator | Herestean Iosif | G.C. | In tintirim | Mirian | 67 de ani | Boale fireasca | G.C. | Berbatu |
| 1860 Decemvrie 8 zile | Decemvrie in 10 | Gavrila Nemesu | Administratore | Gerdalu Malvin | G.C. | In tintirim | Mirian | 13 ani | Boale fireasca | G.C. | berbatu |
| 1860 Decem ber 18 | Decem ber 20 | Gavril Nemes | Administrator | Gerdalu Marine | G.C. | In tintirim | Miriane | 47 | Boale fireasca | G.C. | femeie |
| 1860 decem ber 20 | Decem berie 22 | Gavril Nemesu | Administrator | Gerdalu Samuila | G.C. | In tintirim | Miriane | 67. | Boale fireasca | G.C. | berbatu |
| 1861 in 3 Januarie | Januarie in 5 | Gavril Nemes | Administrator | Herestean Ioan Anisie | G.C. | In tintirim | mirian | 65. | Boale fireasca | G.C. | femeie |
| 1861 in 3 Jan | Jan 5 | Giurgi Poplu | Cantore | Gerdalu Ioane | G.C. | In tinti | miriane | de 7 luni | Boale fireasca | G.C. | femeie |

Note: The columns contain the following information: year, month, day of the death, burial date, name of the deceased, denomination, place of burial, name of the priest, the age of the deceased, cause of death, observations.

## 2.2 SAMPLE STRUCTURE

Since registers covering the project period are kept in the county departments of the National Archive, photographing documents was an important activity. Analysis of the sources, including review of inventories of archival collections was one of the first tasks of the research team, followed by the development of a methodology for selecting micro-areas included in the database. The target was to cover between 5 and 10% of the Transylvanian population from the period 1850–1914. By the end of the project, 7% was included.

The sampling framework was built on several prerequisites (Crăciun, Holom, Popovici, 2015). The first one was continuity of the sources: the parish records must have information for at least 50 consecutive years. Then, the sources had to include the ethnic and denominational composition of Transylvania, ensuring balance with regard to ethnicity and denomination, while taking into account the geographic, ethnocultural, and historical unity/homogeneity of certain regions. An important selection criterion was developed to balance the countryside and settlements with an integrative role, like urban and semi-urban centres. The resulting sample is divided into 12 micro zones which consistently cover almost 7% of the historical population of Transylvania. For an extensive discussion of the sample selection procedures, see Crăciun et al. (2015).

Teams of researchers went to 10 Transylvanian county branches of the National Archive of Romania, gathering the sources and assuring their primary organization (photographing, processing and cataloguing the images). Gathering sources has been an ongoing activity, and there are currently around 500,000 images in the repository of sources. It should be mentioned that no complete and accurate catalogue of these sources in the County Archives Services existed until recently. One of our tasks was to gather all information on existing parish registers and build an electronic catalogue. We estimate that Transylvanian archives hold over 30,000 parish registers, covering seven major denominations, written in five main languages and three alphabets.

Confronting the sources revealed challenges for the team. Registers from the first locality included in the sample (Călăraşi, Cluj county) complied with all of the initial prerequisites, but the registers proved to have very sparse information. Thus, we decided to adapt the structure of the sample by adding the richness of information in the registers to the initial sampling criteria. A decisive argument for our approach was the linkage process performed on the first locality, which encountered numerous difficulties because of the lack of information in the sources.

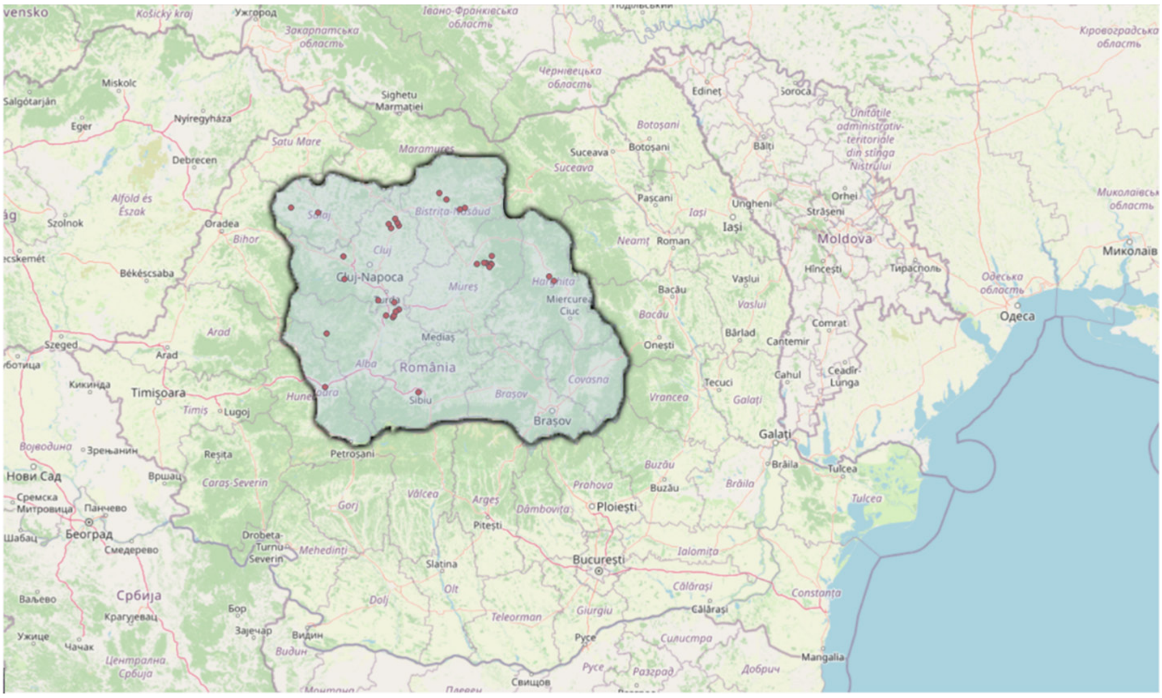
From November 2014 to February 2021, 25 localities were included in HPDT, totalling 165 parish registers. Table 1 provides an overview of the current status of the database containing 141,038 events.

Table 1 *Number of locations and events, included in the HPDT database, February 2021*

| Denomination | BIRTHS | | DEATHS | | MARRIAGES | | BETHROTALS | | **Total** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Locations | Events | Locations | Events | Locations | Events | Locations | Events | Locations | Events |
| Civil registration | | | 1 | 1,337 | | | | | 1 | 1,337 |
| Jewish | 2 | 1,660 | 2 | 397 | 2 | 377 | | | 6 | 2,434 |
| Roman Catholic | 5 | 5,995 | 5 | 6,292 | 4 | 1,927 | | | 14 | 14,214 |
| Greek Catholic | 15 | 21,503 | 14 | 15,974 | 14 | 5,817 | 1 | 188 | 44 | 43,482 |
| Orthodox | 8 | 12,894 | 7 | 8,433 | 7 | 2,630 | 2 | 447 | 24 | 24,404 |
| Lutheran[1] | 1 | 17,241 | 1 | 12,696 | 1 | 5,648 | | | 3 | 35,585 |
| Calvinist (Reformed) | 8 | 8,807 | 8 | 7,476 | 9 | 3,299 | | | 25 | 19,582 |
| Total | 39 | 68,100 | 38 | 52,605 | 37 | 19,698 | 3 | 635 | **117** | **141,038** |

[1] This represents the special case of Sibiu, a town from the southern part the selected area, one of the Transylvanian places inhabited by Saxons. Sibiu served as a case study for a doctoral thesis defended at University of Regensburg in 2021 and the HPDT team gladly accepted to host the data and to provide the database architecture. The data are particularly interesting since Sibiu is one of the big towns of Transylvania, with a population structure more differentiated in terms of education, occupations, intergenerational changes, which will allow complex analyses into the future.

Figure 5 *Map of Transylvania and the localities included in HPDT*



In February 2021 there were 570,036 individuals recorded in HPDT (360,000 in April 2017, at the end of the project). Although the standard period of the database is 1850–1914, due to the necessity of preserving the unity of the archival collection, some localities contain data recorded before 1850 or after 1914. For instance, if the locality preserved records starting from 1780, the project team decided to incorporate the years before 1850 to maintain the integrity of the source and for research purposes. In some situations, the registers did not cover the entire time span of the project or the records did not cover all denominations (the Roman-Catholics and the Reformed started to keep evidence sooner than the Orthodox, and their registers are much better preserved).

## 2.3 GENERAL PRINCIPLES OF DATA ENTRY

The database building process started from the protocol provided by Mandemakers and Dillon (2004). Information was transcribed literally into the database, avoiding any form of abbreviation or standardization. A detailed instruction manual for transcription was created and continuously updated during the process (Bolovan et al., 2019). The data-entry operators were trained with this manual, and they had permanent access to it. However, the variety in information contained in the church registers (multiple languages, different alphabets, and different denominations) required a significant number of adjustments and adaptations in order to conform to international standards of database creation. This meant successive adaptations of the data entry manual and of the database architecture as new particularities of the sources unfolded.

As a general principle of data entry, information from the source was literally transcribed in forms constructed for each type of event (birth, marriage, engagement, death). The data-input forms have various types of fields: text fields, dropdown lists and checkboxes. The data entry operator transcribed information exactly as it appears in the source into text fields, including the errors or misspellings, if any. In this way, the original text of the source was preserved. Such fields included first and last name, nickname, occupation, cause of death etc.

Since the team was reduced in size, the time pressure was high, and some information was repetitive, a decision was made for a preliminary basic standardization of a few fields, in order to avoid redundancy and errors of transcription during the data-entry process. Later standardization would have implied a lot of resources to code and integrate these data into the database structure. Denomination, gender and literacy were standardized for each individual in the event (parents of the baptized child, of the bride and groom, of the deceased, as well as witnesses and godparents), legitimacy for the baptized child, and marital status of the bride and groom or the deceased. Dropdown lists with predetermined values were used for these fields from which the appropriate element must be selected in accordance with the information in the register. For example, the Denomination list includes "Christian", "Jew", "Orthodox", "Greek-Catholic", "Roman-Catholic", "Lutheran".

A similar approach was used for priests and midwives. The same priest could have officiated over 2,000 events (baptisms, marriages and burials); a midwife could have assisted tens of births. It was more efficient to gather all priests present in the sources into one dropdown list, which was updated as new priests appeared in the registers. The same was done with the midwives.

A third type of field used in the data entry forms was checkboxes in which the appropriate value must be ticked, such as stillborn, multiple birth or Julian calendar. The same approach was used to add Observations/Comments, where a ticked box opens a text field to be filled in.

The data entry forms have fields for all the information likely to be recorded in registers, even though the sources differ greatly in the quality and quantity of data. For example, the forms have fields for the occupation of all possible roles (the parents of the baptized, the brides and grooms, the godparents and so on), but many registers lack this information.

## 2.4 THE DATABASE

The original database consisted of tables for each of the vital events found in the parish registers, reflecting our source-oriented approach. Over time, the database and data entry forms evolved as we gained more experience and encountered more complex sources. Features like drop-down lists were added to standardize repetitive information and speed data entry. The database was "normalized" by adding tables for witnesses, godparents, and other participants who appeared in varying numbers. Even when an event is stored in multiple tables, the original document can be reconstituted digitally using keys that link tables to each other. The following section describes phases in the restructuring of the database, which are shown in Figures 6–9 in the Appendix.

### 2.4.1 DEVELOPMENT AND BASIC STRUCTURE

The architecture of the source database needed to combine different parish registers from different denominations into a single structure. It was structured around the main vital events (births, betrothals, marriages, deaths) and included information not only about individuals but also about the event itself. The parish registers from Transylvania contain very diverse information, and each denomination has their own system of recording information with different column headings that changed over time. The database had to accommodate all the different fields from the parish registers so the number of columns for each main table of the database increased to a very large number. However, every type of information was not found in all the registers, and columns would have been empty in more than 70% of records. The information recorded by the priests was sometimes very basic, reduced to the names and possibly a date and place, even though the registers had columns for more information. Tables with more than 200 columns were difficult to manage and had performance issues in retrieving data. Moreover, the empty columns were unnecessarily increasing the size of the database. The solution was to normalize the database and organize the information in linked tables.

In the first version *hpdt_v1* built in 2014, each data entry form describing a main event included all the different fields found in the parish registers from all denominations attached to a table in the database. Information was structured in 17 tables and 475 columns (see Appendix, Figure 6). Although the database followed the source-oriented method, a decision was made to use standardized forms for repetitive information throughout the registers. As we already mentioned, this included denomination, gender, legitimacy and all information regarding priests and midwifes. Priests and midwives were identified by a unique combination of name and place of service.

When the richness of the information in the registers became a main criterion for inclusion in the sample (see Section 2.2), new fields were added in the data entry forms and accommodated in the database as new columns in the underlying tables (for example second occupation for all participants in the events, relation between godparents, information about the spouse of the deceased, ethnicity). Additional data entry forms were added for information about confirmation, converts, and name changes, as this information was found in some of the parish registers. Thus, the second version *hpdt_v2*, implemented by the end of 2015, contained 29 tables and 690 columns (see Appendix, Figure 7). Many of these tables are relational and allow for many-to-many relationships, but they are not included in the graphical interface shown in Figure 7.

In 2016 a third version was necessary, *hpdt_v3.* During the data-entry, new fields emerged, and the database needed to be restructured. The main change was moving the godparents sections from Births and Marriages to a new table. These event tables could accommodate one pair of godparents with information that could entail 15 columns. Further data entry encountered events with multiple pairs of godparents present, ranging from 2 to 5 pairs. Accommodating all of them would have unduly increased the number of columns in the event table, given that a majority of the events recorded had only a single pair. In order to avoid repeating fields the database was further normalized by creating a new table — Godparents — with a corresponding data entry form for the operator. Every pair of godparents was linked by the marriage or birth ID. The complete event was reconstituted in the detailed view of the user interface, so nothing from the original recording of the event is lost. *Hpdt_v3* had 36 tables and 700 columns (see Appendix, Figure 8). So, the detailed view of the data entry form also contains not only the data that has to be entered but also all relevant information from the related tables.

In 2020, normalization was applied to the marriage witnesses, leading to the fourth version, *hpdt_v4*. The data entry form for marriage witnesses allowed up to seven individuals to be recorded, but events including seven witnesses were extremely rare, so many columns remained empty. Another new addition was the table Cause of Death. The cause of death was still recorded in the original language and exactly as it appears in the registers, but all values were written into a dedicated table along with standardized fields on the data entry form. The result was a mix between standardization and retaining the original text of the written source. This change was needed to speed up the process of standardization and to eliminate the data redundancy, a procedure done by the data entry operator. A third addition was the table Death Relatives for relatives present at the Burial event, which used an approach similar to the ones applied to godparents and marriage witnesses to cope with extremely heterogeneous information. With all the newly added tables, *hpdt_v4* has 43 table and 829 columns (see Appendix, Figure 9).

Even though information was separated in different tables in the database architecture, the whole source can be reconstituted digitally to display a copy of the original written record. Everything is linked with the table Sources, in which the original sources are recorded. The user can retrieve the archival code of the registry, the page where the event was mentioned, the languages used and the location. When accessing the information related to the source, the number of records from that particular source is displayed for the user to see how large the source was. The database has a user interface accessible by login.

To summarize, the database is mainly built around four major tables corresponding to the types of vital events recorded in the parish registers. Each table details a single type of religious event: baptism (Births), engagement (Betrothals), marriage (Marriages) and burial (Deaths). Within the tables each row describes a single event. Several other tables include specific persons, such as godparents, witnesses, relatives present at the main events. In addition to the main tables, complementary tables provide values for the dropdown lists that standardize the database. Tables are linked to each other by keys allowing the information to be combined. Tables providing objects for value lists are: Source, Priests, Midwives, Converts, Confirmation, Name change, Denominations, Ethnicities, Countries, Genders, Legitimacies, Dispensations, etc. All events from the main tables are linked to the original written source, and the whole source can be reconstituted digitally.

It is clear that the database is still evolving. If new types of information are found during the transcription process, new fields are added to accommodate the original source. Search filters are provided for researchers in order to find and sort records according to their needs. Everything is interrelated, making the extraction of the necessary information an easy task.

### 2.4.2 INTEGRATED DATABASE: STANDARDIZATION

The second component of the research database is the standard database, which is a copy of the central source database with some additional fields. In the structural metadata of three main tables (Births, Marriages, Deaths) new columns were added for standard names, standard age, and standard places. The original source values were also preserved alongside the standard versions. Values already standardised from the tables for denominations, civil status, legitimacy, literacy, priests, and midwives were not modified. In addition to the tables originating from the source database, several tables have been created to aid in the standardization and record linkage processes. One of the most important tables was the Names table. Standardisation has been applied to the first name, last name and nickname. As we stated before, one of the major characteristics of Transylvania was the use of several languages in the parish registers: Romanian, Hungarian, German, and Latin. Romanian and Hungarian names were processed each in their respective language. Romanian names were assigned for the Orthodox and Greek-Catholics, while the Catholics and Calvinists were associated with Hungarian names. The presence of Hungarian names in the Romanian registers and vice versa were considered acceptable losses. Names like "Catalina" were standardised to "Cătălina" (Romanian) or "Katalin" (Hungarian). Several names with multiple variants, like diminutives, were standardised to the most common one.

In addition to names, other variables were standardized, such as age, location, occupations and causes of death. Age was standardized and divided into several columns for years, months, days. Locations were coded and standardized according to the local administrative unit. We adopted the encoding method that is largely used within the European Union (http://ec.europa.eu/eurostat/web/nuts/local-administrative-units).

The sources included in HPDT provide information on the occupations of individuals with different roles in vital events: mother, father, grandparents, and godparents of the baptised child; groom, bride, parents, godparents, and witnesses in a marriage; the deceased person, parents, and spouse. Following our participation in a workshop on IDS for Population Registers in Lund, Sweden, September 2015, the HPDT core team decided to begin the process of coding Transylvanian historical occupations using the Historical International Standard Classification of Occupation (HISCO) scheme (van Leeuwen, Maas, & Miles, 2002). This scheme, used worldwide, is a necessary step not only to do comparative research but also to standardise the enormous heterogeneity in terms of languages and alphabets in the Transylvanian database. Thus, occupations such as "miller" may appears written in Latin as "molendinarius" or "molitor", as "molnár" in Hungarian, "Müller" in German, or "morar" in Romanian. By using the HISCO scheme, all these linguistic variations are brought together under one single code: 77120.

The coding process of occupations followed several steps: 1. Standardizing the names of the occupations, proofreading and reviewing, the extension of abbreviations; 2. The translation of the occupation names into English; 3. Assigning HISCO codes; 4. Assigning HISCLASS codes; 5. Assigning SOCPO codes. Further, specific historical class or social status schemes, such as HISCLASS (van Leeuwen & Maas, 2011) or SOCPO (Van de Putte & Miles, 2005) were built. For instance, the "miller" with a 77120 code in HISCO, becomes 7 (medium skilled workers) in HISCLASS and 2 (semi-skilled) in the SOCPO scheme. The process of transformation of HISCO codes from HPDT in HISCLASS and SOCPO is undergoing, but these two schemas have already been used in some analyses carried out by members of the Centre for Population Studies despite some misregistration of occupational titles in the parish registers from Transylvania (Holom, Sorescu-Iudean, & Hărăguș, 2018, p. 339).

As part of standardization in HPDT, causes of death have been subjected to linguistic equivalence and coding designed to facilitate the use of information for future demographic research. The coding process involved several steps: 1. Standardization of the causes of death, proofreading, reviewing and completion of abbreviations; 2. English translation (the modern correspondent in English language); 3. Assigning of codes in the International Statistical Classification of Diseases and Related Health Problems-ICD-10.; 4. Assigning of codes in the Historical Causes of Death-HCD (HCD) system (World Health Organization [WHO], 2016; Holom & Hegedűs, 2021).[2]

The process of standardization has been essential not only because registration was in Hungarian, Latin or German, but also because sometimes causes of death appeared in parish registers using colloquial language. For example, "gutaütés", "lovit de gută", "apoplexie", "Schlagfluß" are all terms used to designate stroke. The next step was to code each cause of death following the ICD-10 standard to facilitate future analysis, wider accessibility, and the possibility of comparative studies. Coding into ICD-10 proved to be very problematic. Causes of death mentioned in Transylvanian parish registers from the 18th and 19th centuries are often focused on symptoms (e. g. fever, cough), or are extremely vague (e.g. "natural death", "old age" or "3 days long sickness"), since the vast majority of records were not made by established specialists. Therefore, coding on the basis of ICD-10 created significant difficulties, which we want to overcome by developing a special coding system, the HCD, which is more compatible with data on historical populations.

The Historical Causes of Death system (HCD) is structured in eight categories: 1. Infectious diseases; 2. Chronic and acute non-infectious diseases; 3. Diseases originating in the perinatal period; 4. Diseases

[2] This part of documenting the process of coding of causes of death is a work in progress, and as such the name and the structure of the new system that we intend to develop may have future changes.

related to pregnancy, childbirth and childbed period; 5. Old age-related diseases; 6. Violent deaths; 7. Symptoms, signs and abnormal findings; 8. Ill-defined and unknown causes of mortality. As such the Hungarian term "gutaütés" received the code I64 in ICD-10 and 2 in HCD system. We consider the HCD system suitable for future analysis of mortality in Transylvania, and it can be used by other scholars studying causes of death.

### 2.4.3 INTEGRATED DATABASE: THE LINKAGE PROCESS

At this time, record linkage has only been applied to data from the first locality included in HPDT, namely Călărași, Cluj county. The sample consisted of 2,497 births (baptisms), 1,020 marriages and 2,577 deaths and the individual names, age, locations and birth dates were previously standardized. Only individuals with a main role in the event were included, resulting in a sample of 14,311 individuals from the three types of registers. The fields extracted for record linkage included original names, standard names, location (birth place, residence, wedding place), gender, birth year, event year (when the individual is mentioned), wedding year, death year. First name, last name, sex and role in the event were the variables used to link persons appearing in the registers to unique individuals.

We first linked parents to their (multiple) children in the baptism records. If there were multiple children born from the same parents within a certain time interval the record linkage program assigned the same id number in order to reconstitute the family. Then, we linked baptisms records with the parental marriage records by identifying parents from the baptism records with the bride and groom from a marriage record. Then we linked deaths records as well, identifying parents from the baptism records with their death record. Through a similar process, we identified children in the baptism records as spouses in marriage records and/or as deceased, in deaths records. Difficulties were encountered in every stage, and the linkage process turned into a semi-automatic one.

The software used for record linkage is based on Jaro-Winkler similarities between names. It was developed at the Arctic University of Norway, Tromsø and adapted to the realities of Transylvania. The program uses three levels of the result score: level 1 – score 0.96 (most probable match), level 2 – score 0.90 (probable match), level 3 – score 0.80 (possible match). When using the first level the computer writes the link (gives the matched persons the same ID) automatically into the database. Level 2 and 3 have to be checked by the user before the computer enters the established identification into the database. Only the standard names have been used because the original names had great variability in spelling.

Ethnic and denominational diversity of Transylvania created several challenges. If a person was recorded both in a Romanian and in a Hungarian register with names translated into the respective languages (e.g. "Ioan"/"János"), it was impossible for the automated linkage software to identify he/she as the same person. In this case, only manual linkage could be used because the Romanian and Hungarian versions of a name get low similarity scores.

Naming practices for the female population are also a problem. While Hungarian women kept their maiden name after marriage[3], there was no apparent rule for last names after marriage among Romanian women. Sometimes they were mentioned with their maiden name, other times by their husband's last name, but most of the time they lacked last names altogether. Several variables, like maiden name and role in the event, were constructed in order to adjust the linkage process to this historical reality (Wisselgren, Edvinsson, Berggren, & Larsson, 2014).

However, the biggest problem for accurate linkage was lack of information in the sources, which led to a system of variables created through inference. A semi-automatic linkage was developed for incomplete or heterogeneous information. A series of stored procedures with different query conditions and a Jaro-Winkler function extracted files with possible connections that were checked manually.

The need for homogenization of information for all confessions, missing information, and the construction of multiple supplementary variables were challenges for the linkage process in Transylvania. The results obtained under these circumstances — 73% success rate for linking parents and children, 29% for linking baptism and marriage records and 21% for linking baptism and death records — were a solid argument in favour of the choice to select sources with richer information. The database is still in development, data-

[3] The Hungarian practice of adding the "né" suffix after the husband's name in order to underline the marital status of the wife is well documented (Fercsik, 2010).

entry is a continuing process, and standardization and record linkage are still in an early phase. The next logical steps are to advance with these processes, while accommodating them to Transylvanian realities.

# 3 HPDT AS AN INSTRUMENT FOR RESEARCH

The HPDT longitudinal database was created for research on the population of Transylvania in the 19th and 20th centuries. The database opens new directions of research in areas such as history, demographics, sociology, economy, linguistics, and medical history. The period of time covered by the HPDT represents a crucial era for the study of fertility decline, urbanization, household composition, occupational structure, gender equality. By providing individual-level data, the database allows for statistical analyses with advanced methods on questions that have received very little investigation from a historical perspective. In this section we provide an overview of the main published analyses conducted with data extracted from the HPDT.

An analysis of infant mortality in rural parishes in Transylvania in the second half of the 19th century indicated that almost half of deaths occurred in the first month of life. This disastrous reality was a consequence of the poor conditions of sanitation during the period of pregnancy and the moment of child-birth. Children strong enough to survive the neonatal period mostly died from epidemic diseases, and male infants were the most vulnerable (Coroian, 2017). An investigation of the seasonality of mortality in three Transylvanian settlements between 1887 and 1912 highlighted higher levels during spring and winter, the most vulnerable groups being infants and children (Coroian, 2016).

A working sample of 6,719 adults who died after age 24 has been analysed taking into account both environmental and individual conditions, such as locality type, period, marital status and socioeconomic status. The findings indicated that adults in open localities undergoing industrialization were more prone to premature death, than those living in peripheral, agricultural localities. Between 1850 and 1880, adult mortality was influenced to a greater extent by environmental and epidemiological crisis, but differences were due to economic development and working activities between 1881 and 1914. The main beneficiaries of investments made after 1881 in industry, technology, public sanitation and health care were males (men in agricultural occupations, men employed as semiskilled workers, and unmarried men). Marriage had a protective effect on men, but not on women. After the 1880s, survival prospects improved for both males and females (Holom, Hărăguș, & Bolovan, 2021).

Previous research on age at first marriage in Transylvania focused on nuptial realities in small, isolated villages. With data from the HPDT, we could construct a consistent and coherent sample to consider more explanatory factors. Holom et al. (2018) analysed several factors that influenced the age at first marriage, such as denomination, migration background, and socio-occupational status, as well as broader determinants, such as the time frame and the level of development achieved by settlements under study. Some of the findings were in accord with other areas in Europe, such as the tendency of Roman Catholics to marry later, or the postponement of marriage among men and women with a migrant background. The data indicated that Calvinist women and self-employed men tended to marry later, while both men and women in less developed areas married earlier. The article also took into consideration the interaction between individual factors and broader realities. It found that the level of development of localities was in many cases more important than individual co-variates in determining constraints and opportunities on the marriage market.

Combining HPDT data with information from enrolment records and cadastral registers, Botoș (2019) studied social mobility and the role of education in the Gurghiu Valley, located in the eastern part of Transylvania. Occupational titles were coded into HISCO, and later into HISCLASS 5 (Mandemakers et al., 2013). Her findings indicate that society in the Gurghiu Valley was chiefly agrarian and immobile, and only a small number of people were able to climb the social ladder through education.

The second half of the 19th century witnessed the medicalization of childbirth in Europe and in Transylvania as well. Dumănescu and Eppel (2019) approached the "medicalization" process in both its meanings: the professionalization of healers and the spread of medical care. The paper describes the training of midwives in a society on the periphery of Austro-Hungarian society and the introduction of modern care into one of the most intimate moments of a woman's private life.

Interest in midwifery in Transylvania came from the abundance of midwives in the parish registers included into the Historical Population Database of Transylvania. Over a period of less than 50 years, thousands of women performed deliveries in villages and had their names written in the "midwife" column in the registers. This discovery raised questions: Who were these midwives? Were they really midwives? Were they trained, skilled midwives or simply handy women who helped deliver their neighbours' babies? The preliminary results of Dumănescu and Bolovan (2021a) confirm that the medicalization of childbirth at the turn of the 20th century was not nearly as widespread as one might expect from official statistics. Over 80% of the women reported in our sample were handywomen, rather than certified midwives. One could also conclude that the medicalization in all its aspects, including childbirth, went hand in hand with the processes of modernisation and industrialisation.

A study concerning women and their married names in Transylvania in the second half of 19th century Dumănescu and Bolovan (2021b), based on a sample which included 29,000 baptisms, 3,982 weddings, and 6,592 events of death, revealed that a marriage contract was not automatically followed by a change in name and that a married woman was still recognized by her maiden name in subsequent documents.

A study of so-called “necronymic” names (Mârza, 2017) aimed to reconstruct certain relational patterns over several generations of the same family. The article begins from the assumption that the naming of children was not random. This implies that the recurrence of certain names across three or four generations could be an indication of the type and the quality of the constructed links in a family. The practice of “necronymic naming” (naming a child after a deceased sibling) was considered a method of strengthening and perpetuating the fabric of the family. Working on the data from two localities in the HPDT, this article highlighted the limitations of the vital registers for accurately reconstituting all families on both the maternal and the paternal lines. Even when the standardization of HPDT will be complete, the lack of certain essential data (such as the mother's last name in many birth registrations) sometimes limits the reconstitution of families to only one or two generations.

# 4 CONCLUSIONS

From the beginning, HPDT was intended to become an instrument dedicated to researchers with an interest in the past of Transylvania. Designing and building a database, especially when using a source-oriented approach was a difficult task, and HPDT had to cope with complex sources providing heterogeneous information. The intermediate versions of HPDT show how it has continuously evolved and adapted to meet these challenges, ensuring its value for a wide range of research questions and methodologies.

The source-oriented approach preserved the original documents in their entirety, but stored them in an organized form. Standardization of information was the next step for integrating Transylvanian data into international research. Standardization occurred in two steps. The initial step standardized and codified repetitive data, such as denomination, ethnicity, marital status, as well as priests and midwives who appeared frequently in the sources, to eliminate redundancy and reduce transcription errors during the data-entry process. A second round of standardization was applied at the end of data input for selected communities. This time all names, occupations, locations, and causes of death were standardized and codified according to international standards for historical databases. Standardization allowed a partly automated linkage process, using software based on the Jaro-Winkler distance for similarities. However, the lack of uniformity in data from the various Transylvanian denominations is still an impediment to increasing the automation of the linkage process, which will need to be addressed in the future.

The real impact of HPDT can be best assessed from the diversity of studies already published or in progress, which reflects its increased importance for historical, demographic, sociological, and anthropological research. HPDT is continuously adapting to meet the needs of researchers applying micro-level quantitative perspectives to the study of the historical past in Transylvania. Their studies will be better integrated into universal knowledge of the past, because alignment with international standards of databases for historical data makes them suitable for comparisons across national borders.

## REFERENCES

Alter, G., & Mandemakers, K. (2014). The Intermediate Data Structure (IDS) for longitudinal historical microdata, version 4. *Historical Life Course Studies, 1*, 1–26. doi: 10.51964/hlcs9290

Bolovan, I., Crăciun, B., Covaci, D., Dumănescu, L., Holom, E.-C., Mârza, D., & Lumezeanu, A. C. (2019). Historical Population Database of Transylvania. A database manual. *Studia Universitatis Babeș-Bolyai Digitalia*, *64*(1), 9–84. doi: 10.24193/subbdigitalia.2019.1.1

Botoș, R. (2019). Education as a vehicle for social mobility in the 19th century in Transylvania. A comparative view on Romanians and Hungarians in the Gurghiu valley. *Romanian Journal of Population Studies*, *13*(1), 29–46. doi: 10.24193/RJPS.2019.1.02

Coroian, I. G. (2016). The seasonality of mortality in three Transylvanian settlements in the second half of the 19th century. *Romanian Journal of Population Studies*, *10*(1), 19–35.

Coroian, I. G. (2017). Infant mortality in rural Transylvania: A case study on four parishes in the second half of the 19th century. *The Romanian Journal of Modern History*, *8*(1–2), 5–18.

Crăciun, B., Holom, E. C., & Popovici, V. (2015). Historical Population Database on Transylvania: Methodology employed in the selection of settlements and micro zones of interest. *Romanian Journal of Population Studies, 9*(1), 17–31.

Dumănescu, L., & Eppel, M. (2019). The politics of birth in a composite state: Midwives in Transylvania (19th–20th century). *Romanian Journal of Population Studies*, *13*(1), 7–27. doi: 10.24193/RJPS.2019.1.01

Dumănescu, L., & Bolovan, I. (2021a). Medicalisation of birth in Transylvania in the second half of the 19th century. A subject to be investigated. *Historical Life Course Studies, 10*(3), 91–95. doi: 10.51964/hlcs9574

Dumănescu, L., & Bolovan, I. (2021b). 'From the cradle to the grave I am my father's daughter!' Women and their married names in Transylvania in the second half of 19th century**.** *The History of the Family, 26*(3), 466–481. doi: 10.1080/1081602X.2021.1933126

Fercsik, E. **(**2010). The traditional and modern forms of Hungarian female matrimonial names. In M. G. Arcamone, D. Bremer, D. De Camilli & B. Porcelli (Eds.), *Atti del XXII Congresso Internazionale di Scienze Onomastiche Pisa, 28 agosto – 4 settembre 2005* (Vol. IV, Antroponomastica) (pp. 131–140*)*. Pisa: Edizioni Ets. Retrieved from https://mnytud.arts.unideb.hu/nevtan/informaciok/pisa/fe-a.pdf

Holom, E. C., Hărăguș, M., & Bolovan, I. (2021). Socioeconomic and marital status inequalities in longevity: Adult mortality in Transylvania, 1850–1914. *Journal of Interdisciplinary History*, 51(4), 533–564. doi: 10.1162/jinh_a_01627

Holom, E. C., & Hegedűs, N. (2021, April). From ICD-10 to a new nosological classification of causes of death in Transylvania, 1850 and 1920. Talk presented at the *Local Population Studies Society Spring Conference*, Local Population Studies Society and the Southampton Centre for Nineteenth-Century Study, University of Southampton, U.K.

Holom, E. C., Sorescu-Iudean, O., & Hărăguș, M. (2018). Beyond the visible pattern: Historical particularities, development, and age at first marriage in Transylvania, 1850–1914. *The History of the Family*, *23*(2), 329–358. doi: 10.1080/1081602X.2018.1433702

Lumezeanu, A.-C. (2019). *Digital infrastructure for social history. Building historical databases* (Doctoral dissertation). Babeș-Bolyai University, Cluj-Napoca.

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

Mandemakers, K., Muurling, S., Maas, I., Van de Putte, B., Zijdeman, R. L., Lambert, P. S., van Leeuwen, M. H. D., van Poppel, F. W. A., & Miles, A. (2013). *HSN standardized, HISCO-coded and classified occupational titles, release 2013.01*. Amsterdam: IISG.

Mârza, D. (2017). Patterns in family relationships in 19th century Transylvania: Data from the Historical Population Database of Transylvania. *Transylvanian Review*, *26*(4), 63–70.

Van de Putte, B., & Miles, A. (2005). A social classification scheme for historical occupational data. *Historical Methods. A Journal of Quantitative and Interdisciplinary History*, *38*(2), 61–94. doi: 10.3200/HMTS.38.2.61-94

van Leeuwen, M. H. D., Maas, I., & Miles, A. (2002). *HISCO. Historical International Standard Classification of Occupations*. Leuven: Leuven University Press.

van Leeuwen, M. H. D., & Maas, I. (2011). *HISCLASS. A Historical International Social Class Scheme*. Leuven: Leuven University Press.

Wisselgren, M. J., Edvinsson, S., Berggren, M., & Larsson, M. (2014). Testing methods of record linkage on Swedish censuses. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 47*(3), 138–151. doi: 10.1080/01615440.2014.913967

World Health Organization [WHO]. (2016). *ICD-10: International statistical classification of diseases and related health problems, 10th revision, 5th edition* (Vol. 1–3). Retrieved from https://apps.who.int/iris/handle/10665/246208

## APPENDIX — MODELS OF THE DATABASE

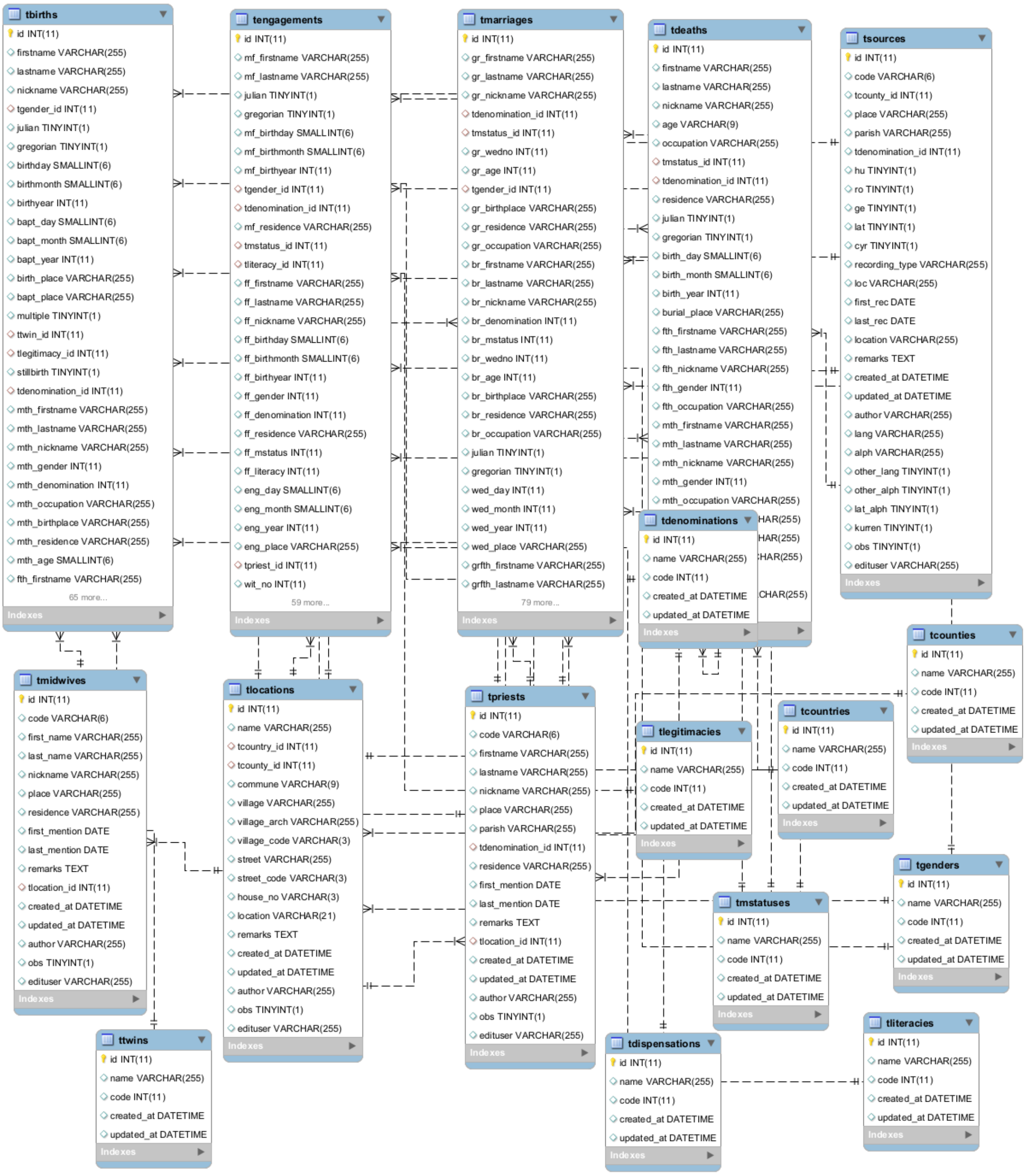


Figure 7 *Model hpdt_v2*

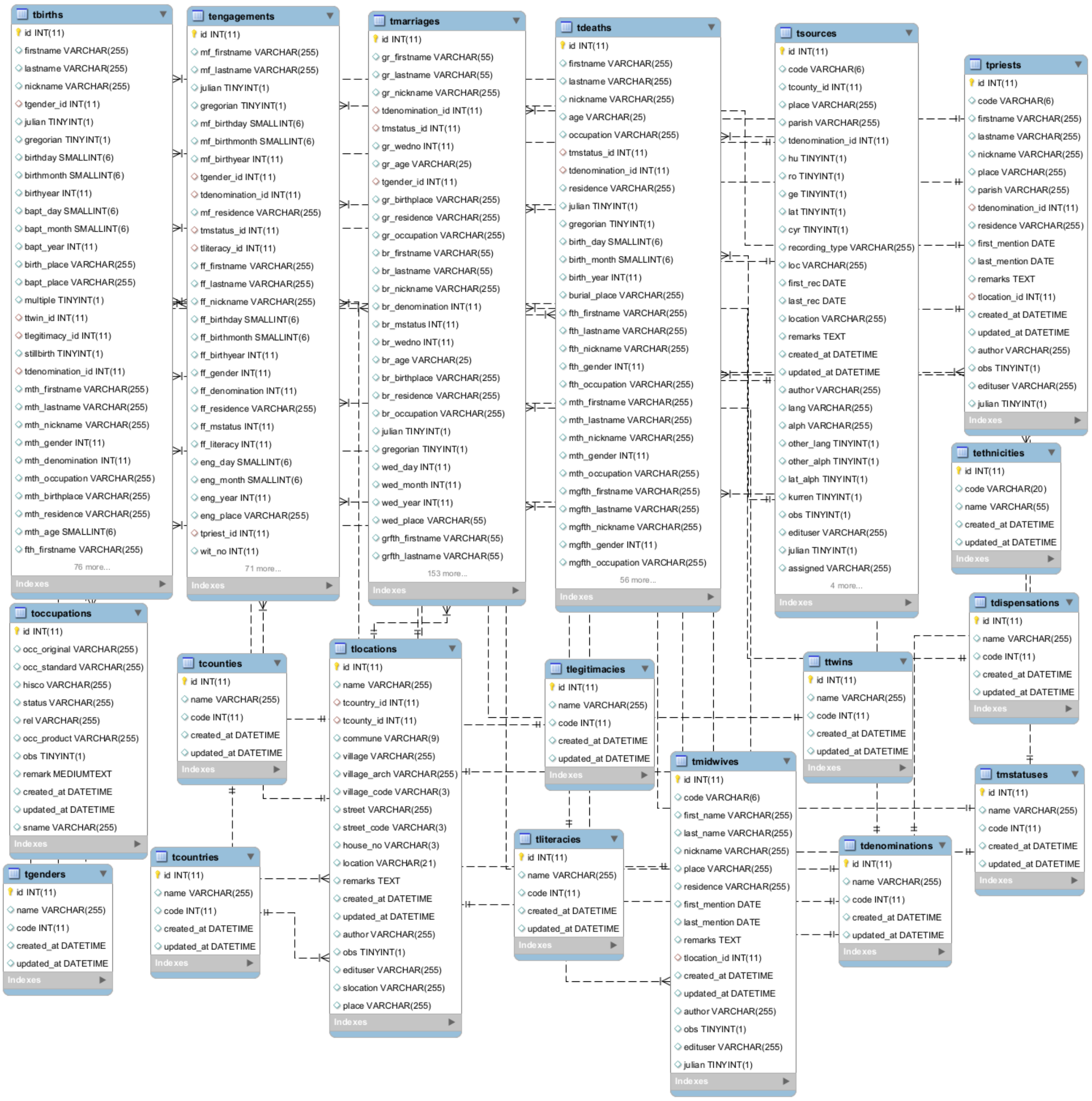


Figure 8 *Model hpdt_v3*

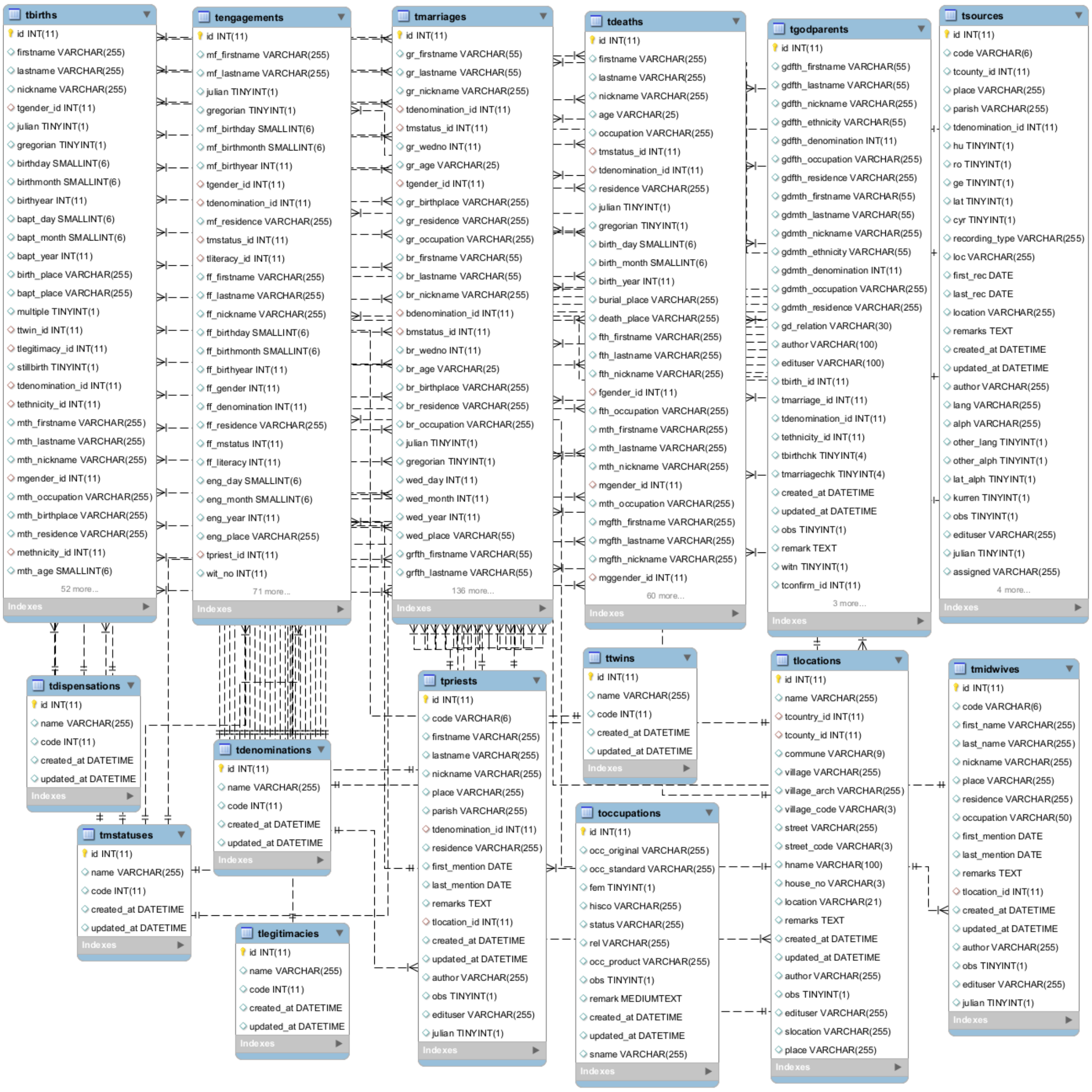
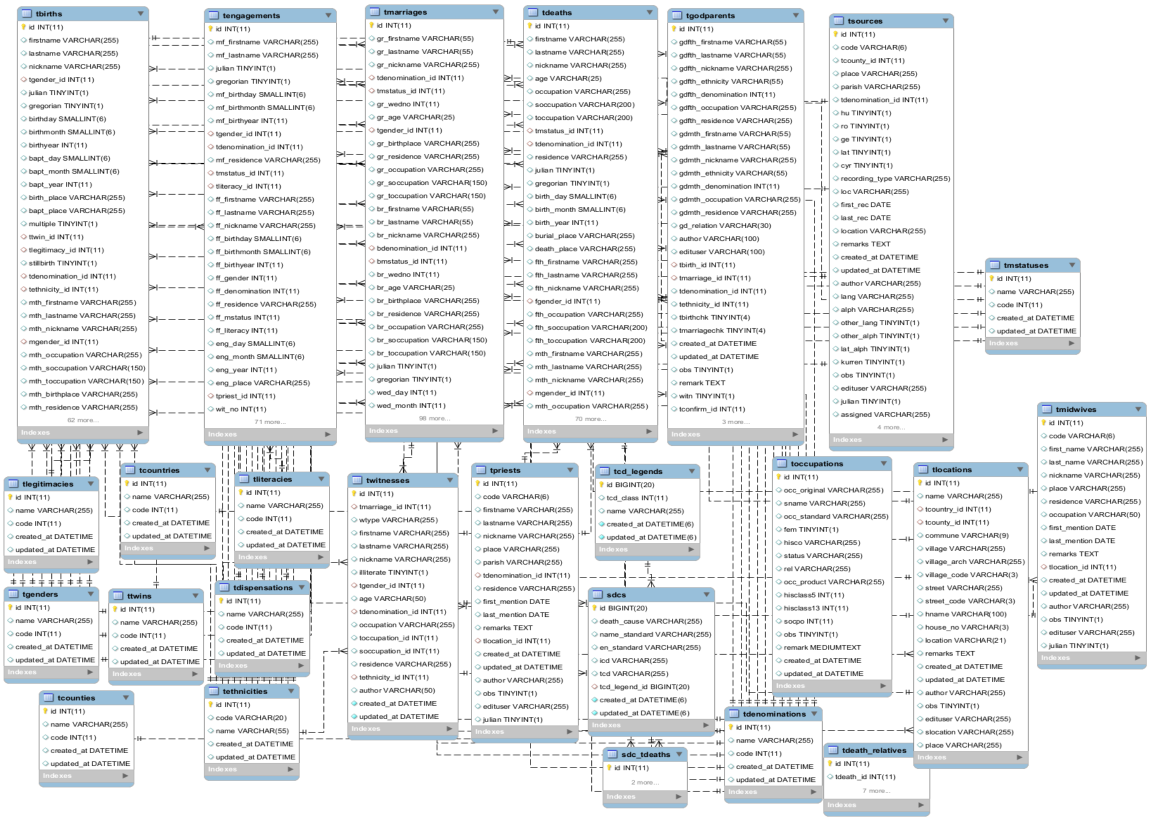


Figure 9 *Model hpdt_v4*


tbirths
id INT(11)
firstname VARCHAR(255)
lastname VARCHAR(255)
nickname VARCHAR(255)
tgender_id INT(11)
julian TINYINT(1)
gregorian TINYINT(1)
birthday SMALLINT(6)
birthmonth SMALLINT(6)
birthyear INT(11)
bapt_day SMALLINT(6)
bapt_month SMALLINT(6)
bapt_year INT(11)
birth_place VARCHAR(255)
bapt_place VARCHAR(255)
multiple TINYINT(1)
ttwin_id INT(11)
tlegitimacy_id INT(11)
stillbirth TINYINT(1)
tdenomination_id INT(11)
tethnicity_id INT(11)
mth_firstname VARCHAR(255)
mth_lastname VARCHAR(255)
mth_nickname VARCHAR(255)
mgender_id INT(11)
mth_occupation VARCHAR(255)
mth_soccupation VARCHAR(150)
mth_toccupation VARCHAR(150)
mth_birthplace VARCHAR(255)
mth_residence VARCHAR(255)
62 more...
Indexes
tengagements
id INT(11)
mf_firstname VARCHAR(255)
mf_lastname VARCHAR(255)
julian TINYINT(1)
gregorian TINYINT(1)
mf_birthday SMALLINT(6)
mf_birthmonth SMALLINT(6)
mf_birthyear INT(11)
tgender_id INT(11)
tdenomination_id INT(11)
mf_residence VARCHAR(255)
tmstatus_id INT(11)
tliteracy_id INT(11)
ff_firstname VARCHAR(255)
ff_lastname VARCHAR(255)
ff_nickname VARCHAR(255)
ff_birthday SMALLINT(6)
ff_birthmonth SMALLINT(6)
ff_birthyear INT(11)
ff_gender INT(11)
ff_denomination INT(11)
ff_residence VARCHAR(255)
ff_mstatus INT(11)
ff_literacy INT(11)
eng_day SMALLINT(6)
eng_month SMALLINT(6)
eng_year INT(11)
eng_place VARCHAR(255)
tpriest_id INT(11)
wit_no INT(11)
71 more...
Indexes
tmarriages
id INT(11)
gr_firstname VARCHAR(55)
gr_lastname VARCHAR(55)
gr_nickname VARCHAR(255)
tdenomination_id INT(11)
tmstatus_id INT(11)
gr_wedno INT(11)
gr_age VARCHAR(25)
tgender_id INT(11)
gr_birthplace VARCHAR(255)
gr_residence VARCHAR(255)
gr_occupation VARCHAR(255)
gr_soccupation VARCHAR(150)
gr_toccupation VARCHAR(150)
br_firstname VARCHAR(55)
br_lastname VARCHAR(55)
br_nickname VARCHAR(255)
bdenomination_id INT(11)
bmstatus_id INT(11)
br_wedno INT(11)
br_age VARCHAR(25)
br_birthplace VARCHAR(255)
br_residence VARCHAR(255)
br_occupation VARCHAR(255)
br_soccupation VARCHAR(150)
br_toccupation VARCHAR(150)
julian TINYINT(1)
gregorian TINYINT(1)
wed_day INT(11)
wed_month INT(11)
98 more...
Indexes
tdeaths
id INT(11)
firstname VARCHAR(255)
lastname VARCHAR(255)
nickname VARCHAR(255)
age VARCHAR(25)
occupation VARCHAR(255)
soccupation VARCHAR(200)
toccupation VARCHAR(200)
tmstatus_id INT(11)
tdenomination_id INT(11)
residence VARCHAR(255)
julian TINYINT(1)
gregorian TINYINT(1)
birth_day SMALLINT(6)
birth_month SMALLINT(6)
birth_year INT(11)
burial_place VARCHAR(255)
death_place VARCHAR(255)
fth_firstname VARCHAR(255)
fth_lastname VARCHAR(255)
fth_nickname VARCHAR(255)
fgender_id INT(11)
fth_occupation VARCHAR(255)
fth_soccupation VARCHAR(200)
fth_toccupation VARCHAR(200)
mth_firstname VARCHAR(255)
mth_lastname VARCHAR(255)
mth_nickname VARCHAR(255)
mgender_id INT(11)
mth_occupation VARCHAR(255)
70 more...
Indexes
tgodparents
id INT(11)
gdfth_firstname VARCHAR(55)
gdfth_lastname VARCHAR(55)
gdfth_nickname VARCHAR(255)
gdfth_ethnicity VARCHAR(55)
gdfth_denomination INT(11)
gdfth_occupation VARCHAR(255)
gdfth_residence VARCHAR(255)
gdmth_firstname VARCHAR(55)
gdmth_lastname VARCHAR(55)
gdmth_nickname VARCHAR(255)
gdmth_ethnicity VARCHAR(55)
gdmth_denomination INT(11)
gdmth_occupation VARCHAR(255)
gdmth_residence VARCHAR(255)
gd_relation VARCHAR(30)
author VARCHAR(100)
edituser VARCHAR(100)
tbirth_id INT(11)
tmarriage_id INT(11)
tdenomination_id INT(11)
tethnicity_id INT(11)
tbirthchk TINYINT(4)
tmarriagechk TINYINT(4)
created_at DATETIME
updated_at DATETIME
obs TINYINT(1)
remark TEXT
witn TINYINT(1)
tconfirm_id INT(11)
3 more...
Indexes
tsources
id INT(11)
code VARCHAR(6)
tcounty_id INT(11)
place VARCHAR(255)
parish VARCHAR(255)
tdenomination_id INT(11)
hu TINYINT(1)
ro TINYINT(1)
ge TINYINT(1)
lat TINYINT(1)
cyr TINYINT(1)
recording_type VARCHAR(255)
loc VARCHAR(255)
first_rec DATE
last_rec DATE
location VARCHAR(255)
remarks TEXT
created_at DATETIME
updated_at DATETIME
author VARCHAR(255)
lang VARCHAR(255)
alph VARCHAR(255)
other_lang TINYINT(1)
other_alph TINYINT(1)
lat_alph TINYINT(1)
kurren TINYINT(1)
obs TINYINT(1)
edituser VARCHAR(255)
julian TINYINT(1)
assigned VARCHAR(255)
4 more...
Indexes
tmstatuses
id INT(11)
name VARCHAR(255)
code INT(11)
created_at DATETIME
updated_at DATETIME
Indexes
tmidwives
id INT(11)
code VARCHAR(6)
first_name VARCHAR(255)
last_name VARCHAR(255)
nickname VARCHAR(255)
place VARCHAR(255)
residence VARCHAR(255)
occupation VARCHAR(50)
first_mention DATE
last_mention DATE
remarks TEXT
tlocation_id INT(11)
created_at DATETIME
updated_at DATETIME
author VARCHAR(255)
obs TINYINT(1)
edituser VARCHAR(255)
julian TINYINT(1)
Indexes
tlegitimacies
id INT(11)
name VARCHAR(255)
code INT(11)
created_at DATETIME
updated_at DATETIME
Indexes
tcountries
id INT(11)
name VARCHAR(255)
code INT(11)
created_at DATETIME
updated_at DATETIME
Indexes
tliteracies
id INT(11)
name VARCHAR(255)
code INT(11)
created_at DATETIME
updated_at DATETIME
Indexes
twitnesses
id INT(11)
tmarriage_id INT(11)
wtype VARCHAR(255)
firstname VARCHAR(255)
lastname VARCHAR(255)
nickname VARCHAR(255)
illiterate TINYINT(1)
tgender_id INT(11)
age VARCHAR(50)
tdenomination_id INT(11)
occupation VARCHAR(255)
toccupation_id INT(11)
soccupation_id INT(11)
residence VARCHAR(255)
tethnicity_id INT(11)
author VARCHAR(50)
created_at DATETIME
updated_at DATETIME
Indexes
tpriests
id INT(11)
code VARCHAR(6)
firstname VARCHAR(255)
lastname VARCHAR(255)
nickname VARCHAR(255)
place VARCHAR(255)
parish VARCHAR(255)
residence VARCHAR(255)
tdenomination_id INT(11)
first_mention DATE
last_mention DATE
remarks TEXT
tlocation_id INT(11)
created_at DATETIME
updated_at DATETIME
author VARCHAR(255)
obs TINYINT(1)
edituser VARCHAR(255)
julian TINYINT(1)
Indexes
tcd_legends
id BIGINT(20)
tcd_class INT(11)
name VARCHAR(255)
created_at DATETIME(6)
updated_at DATETIME(6)
Indexes
toccupations
id INT(11)
occ_original VARCHAR(255)
sname VARCHAR(255)
occ_standard VARCHAR(255)
fem TINYINT(1)
hisco VARCHAR(255)
status VARCHAR(255)
rel VARCHAR(255)
occ_product VARCHAR(255)
hisclass5 INT(11)
hisclass13 INT(11)
socpo INT(11)
obs TINYINT(1)
remark MEDIUMTEXT
created_at DATETIME
updated_at DATETIME
Indexes
tlocations
id INT(11)
name VARCHAR(255)
tcountry_id INT(11)
tcounty_id INT(11)
commune VARCHAR(9)
village VARCHAR(255)
village_arch VARCHAR(255)
village_code VARCHAR(3)
street VARCHAR(255)
street_code VARCHAR(3)
hname VARCHAR(100)
house_no VARCHAR(3)
location VARCHAR(21)
remarks TEXT
created_at DATETIME
updated_at DATETIME
author VARCHAR(255)
obs TINYINT(1)
edituser VARCHAR(255)
slocation VARCHAR(255)
place VARCHAR(255)
Indexes
tgenders
id INT(11)
name VARCHAR(255)
code INT(11)
created_at DATETIME
updated_at DATETIME
Indexes
ttwins
id INT(11)
name VARCHAR(255)
code INT(11)
created_at DATETIME
updated_at DATETIME
Indexes
tdispensations
id INT(11)
name VARCHAR(255)
code INT(11)
created_at DATETIME
updated_at DATETIME
Indexes
sdcs
id BIGINT(20)
death_cause VARCHAR(255)
name_standard VARCHAR(255)
en_standard VARCHAR(255)
icd VARCHAR(255)
tcd VARCHAR(255)
tcd_legend_id BIGINT(20)
created_at DATETIME(6)
updated_at DATETIME(6)
Indexes
tcounties
id INT(11)
name VARCHAR(255)
code INT(11)
created_at DATETIME
updated_at DATETIME
Indexes
tethnicities
id INT(11)
code VARCHAR(20)
name VARCHAR(55)
created_at DATETIME
updated_at DATETIME
Indexes
tdenominations
id INT(11)
name VARCHAR(255)
code INT(11)
created_at DATETIME
updated_at DATETIME
Indexes
sdc_tdeaths
id INT(11)
2 more...
Indexes
tdeath_relatives
id INT(11)
tdeath_id INT(11)
7 more...
Indexes