


# *link-* **LIVES**

## **Release 2 Guide**

Barbara Revuelta-Eugercios, Olivia Robinson, Nicolai Rask  
Mathiesen, Asbjørn Romvig Thomsen, Signe Trolle  
Gronemann, Lise Bødtker Sunde, Ida Dorph Nørskov, Anna  
Lodberg Sparre, Tobias Kallehauge and Anne Løkke



December 2025



Rigsarkivet



UNIVERSITY OF  
COPENHAGEN



KØBENHAVNS  
STADSARKIV

CARLSBERGFONDET

 **nnovationsfonden**

# Link-Lives Release 2 Guide

Barbara Revuelta-Eugercios

Olivia Robinson

Nicolai Rask Mathiesen

Asbjørn Romvig Thomsen

Signe Trolle Gronemann

Lise Bødtker Sunde

Ida Dorph Nørskov

Anna Lodberg Sparre

Tobias Kallehauge

Anne Løkke

December 2025

# Contributors to Link-Lives data and documentation

## Institutions

University of Copenhagen: SAXO Institute, University of Copenhagen: Novo Nordisk Center for Protein Research, Copenhagen City Archives (*Københavns Stadsarkiv*), The National Archives (*Rigsarkivet*).

## Funding

Innovation Fund Denmark 2019-2022, Grant number 8088-00034A. Carlsberg-fondet 2019-2025, grant number CF18-1116

## Principal investigators

Barbara Revuelta-Eugercios and Anne Løkke

## Members of the Steering Committee

Anne Løkke, Mads Neuhard, Ole Magnus Andersen, Kirsten Villadsen Kristmar, Søren Brunak

## Members of Advisory Board

Angelique Janssen (Radboud University Nijmegen), Chris Dibben (University of Edinburgh), Elisabeth Engberg (Umeå University), Hilde Sommerseth (Tromsø University), Kirsten Sanders (Danske Slægtforskere), Nikolai Donitzky (Ancestry)

## Members of the working group

Anne Løkke, Barbara Revuelta-Eugercios, Helene Castenbrandt, Katrine Tovgaard Olsen, Markus Schunck, Signe Trolle Gronemann.

## Data-scientist and it-development specialists

Anders Enghøj, Bo Henriksen, Jakob Humlegaard, Nicolai Rask Mathiesen, Niels Abildgaard, Tobias Kallehauge, Signe Trolle Gronemann.

## Researchers

Asbjørn Romvig Thomsen, Birgit Eggert, Helene Castenbrandt, Olivia Robinson

## PhD students

Anna Lodberg Sparre, Louise Villefrance Perner, Line Hjort-Mouritzen, Mads Villefrance Perner, Samantha Nordholt Aagaard, Roc Reguant.

## **Research assistants**

Atlanta Majka Nørup Young, Anna Kristiane Mortensen, Anna Lodberg Sparre, Ida Dorph Nørskov, Lise Bødtker Sunde, Maria Nathalia Vinter, Line Hjort-Mouritzen, Lasse Kirk Ladekjær Jacobsen, Mads Villefrance Perner, Samantha Nordholt Aagaard.

## **Consulting archivists and archive managers**

Allan Vestergaard, Anders Sode-Pedersen, Anne Sofie Fink, Bente Vestergaard, Christian Babiarz Madsen, Helle Damgaard Andersen, Jan Dalsten, Jeppe Christensen, Lisette Rønsig Larsen, Louise Villefrance Perner, Mikkel Eide Eriksen, Peter Brix, Per Seesko-Tønnesen

## **Student assistants**

Aima Hussain, Anna Lodberg Sparre, Evangelia Miariti, Lise Bødtker Sunde, Sif Adriana Stavnstrup

## **Interns and volunteers**

Aima Hussain, Anne Sofie Boye Hansen, Christina Vibeke Holck-Clausen, Clara Dalsgaard Hansen, Ellen Munk Ebbesen, Frida Helene Beck-Larsen, Hannah Emilie Roguelin de Place, Henrik Iversen, Jesper Mühlback Hansen, Ida Asp, Laura Riis Nielsen, Sofie Vestergaard, Maria Stockholm Van, Hanne Jensen, Sif Adriana Stavnstrup.

# Contents

|  |           |
|--|-----------|
| <b>List of Figures</b>   | <b>9</b>  |
| <b>List of Tables</b>  | <b>11</b> |
| <b>1 Preface</b>   | <b>13</b> |
| <b>2 Introduction</b>  | <b>15</b> |
| 2.1 Purpose and overview of the guide . . . . .  | 15        |
| 2.2 Overview of release: datasets and availability . . . . .                                       | 16        |
| 2.3 Types of files and where to find them . . . . .  | 16        |
| 2.3.1 Source files . . . . .   | 17        |
| 2.3.2 Link files . . . . .   | 18        |
| 2.3.3 Auxiliary data . . . . .   | 19        |
| 2.3.4 Data formats . . . . .   | 19        |
| 2.4 What's new in release 2? . . . . .   | 19        |
| 2.5 A searchable overview of the release: <a href="http://linklives.dk">linklives.dk</a> . . . . . | 20        |
| 2.6 How to cite . . . . .  | 20        |
| 2.7 License . . . . .  | 21        |
| 2.7.1 All sources . . . . .  | 21        |
| 2.7.2 The fully transcribed censuses 1787-1901 . . . . .   | 21        |
| 2.7.3 Transcription of parish registers . . . . .  | 21        |
| 2.7.4 The Copenhagen Burial Register 1861-1911 . . . . .   | 21        |
| 2.7.5 Links, life-courses and the Link-Lives harmonized versions<br>of the above data . . . . .    | 21        |
| 2.8 Conventions used in this guide . . . . .   | 21        |
| 2.9 Contact . . . . .  | 22        |
| <b>3 Historical sources</b>  | <b>23</b> |
| 3.1 Censuses . . . . .   | 23        |
| 3.1.1 Availability of images . . . . .   | 23        |
| 3.1.2 History . . . . .  | 23        |
| 3.1.3 Who was registered where . . . . .   | 26        |
| 3.1.4 Geographical coverage . . . . .  | 26        |
| 3.1.5 Legal basis and instructions for the registration . . . . .                                  | 27        |
| 3.1.6 The layout of the pre-printed forms and information recorded . . . . .                       | 27        |
| 3.1.7 Other variables . . . . .  | 29        |
| 3.1.8 Published, aggregate census results and analyses . . . . .                                   | 30        |
| 3.1.9 Further reading . . . . .  | 30        |
| 3.2 Parish Registers . . . . .   | 31        |
| 3.2.1 Availability of images . . . . .   | 31        |
| 3.2.2 History . . . . .  | 32        |
| 3.2.3 Geographical coverage . . . . .  | 33        |
| 3.2.4 Information recorded . . . . .   | 33        |
| 3.2.5 Published, aggregate vital statistics . . . . .  | 36        |
| 3.2.6 Further reading . . . . .  | 36        |
| 3.3 Copenhagen Burial Register . . . . .   | 37        |

|          |  |           |
|----------|--|-----------|
| 3.3.1    | Availability of images . . . . .                   | 37        |
| 3.3.2    | Coverage . . . . .                                 | 38        |
| 3.3.3    | Variables . . . . .                                | 38        |
| <b>4</b> | <b>Transcribed datasets</b>                        | <b>40</b> |
| 4.1      | Censuses . . . . .                                 | 43        |
| 4.1.1    | Background to the transcription . . . . .          | 43        |
| 4.1.2    | Transcription principles . . . . .                 | 46        |
| 4.1.3    | Known data issues . . . . .                        | 49        |
| 4.2      | Parish registers . . . . .                         | 50        |
| 4.2.1    | Background to the transcription . . . . .          | 50        |
| 4.2.2    | Transcription principles . . . . .                 | 51        |
| 4.2.3    | Known data issues . . . . .                        | 54        |
| 4.3      | Copenhagen Burial Register . . . . .               | 55        |
| 4.3.1    | Transcription principles . . . . .                 | 56        |
| 4.3.2    | Known data issues . . . . .                        | 60        |
| <b>5</b> | <b>Link-Lives data harmonisation</b>               | <b>61</b> |
| 5.1      | Establishing source IDs . . . . .                  | 61        |
| 5.2      | Identifying person appearances . . . . .           | 61        |
| 5.2.1    | Censuses and Copenhagen Burial Registers . . . . . | 63        |
| 5.2.2    | Parish registers . . . . .                         | 63        |
| 5.3      | Event_type . . . . .                               | 64        |
| 5.4      | Role . . . . .                                     | 65        |
| 5.5      | Event date variables . . . . .                     | 65        |
| 5.5.1    | Census . . . . .                                   | 65        |
| 5.5.2    | Parish registers . . . . .                         | 66        |
| 5.5.3    | Copenhagen burial register . . . . .               | 67        |
| 5.6      | Age/birthdate . . . . .                            | 67        |
| 5.6.1    | Census . . . . .                                   | 67        |
| 5.6.2    | Parish registers . . . . .                         | 67        |
| 5.6.3    | Copenhagen burial registers . . . . .              | 67        |
| 5.7      | Event_place . . . . .                              | 68        |
| 5.7.1    | Census . . . . .                                   | 68        |
| 5.7.2    | Parish registers . . . . .                         | 69        |
| 5.7.3    | Copenhagen Burial Register . . . . .               | 69        |
| 5.8      | Birth places . . . . .                             | 69        |
| 5.8.1    | Census . . . . .                                   | 69        |
| 5.8.2    | Parish registers . . . . .                         | 70        |
| 5.8.3    | Copenhagen burials . . . . .                       | 70        |
| 5.9      | Names . . . . .                                    | 70        |
| 5.9.1    | Census . . . . .                                   | 70        |
| 5.9.1.1  | Cleaning . . . . .                                 | 71        |
| 5.9.1.2  | Standardisation . . . . .                          | 71        |
| 5.9.1.3  | Classification . . . . .                           | 71        |
| 5.9.1.4  | Resulting variables . . . . .                      | 72        |
| 5.9.2    | Parish registers . . . . .                         | 72        |

|          |  |           |
|----------|--|-----------|
| 5.9.3    | Copenhagen burial registers                              | 73        |
| 5.10     | Sex  | 73        |
| 5.10.1   | Census   | 73        |
| 5.10.2   | Parish registers   | 74        |
| 5.10.3   | Copenhagen burial registers                              | 74        |
| 5.11     | Civil status   | 74        |
| 5.11.1   | Census   | 74        |
| 5.11.2   | Parish registers   | 75        |
| 5.11.3   | Copenhagen burial registers                              | 75        |
| <b>6</b> | <b>Linking methods</b>                                   | <b>76</b> |
| 6.1      | Computer-assisted domain-expert record linkage           | 76        |
| 6.1.1    | Purpose  | 81        |
| 6.1.2    | What we link? Linking scope                              | 81        |
| 6.1.2.1  | Sequence   | 81        |
| 6.1.2.2  | Target units and sampling strategy                       | 81        |
| 6.1.2.3  | Sampling of individuals within linking units             | 81        |
| 6.1.3    | How we link? Specific choices in Linking method          | 82        |
| 6.1.3.1  | Data allowed for linking                                 | 82        |
| 6.1.3.2  | Choice of variables                                      | 82        |
| 6.1.3.3  | Software assistance                                      | 83        |
| 6.1.3.4  | Direction of linking                                     | 85        |
| 6.1.3.5  | Iterations   | 86        |
| 6.1.4    | Who links? Choice of linkers                             | 86        |
| 6.1.4.1  | Number of linkers and adjudication methods               | 86        |
| 6.1.4.2  | Choice and training of linkers                           | 87        |
| 6.1.5    | Results  | 87        |
| 6.1.5.1  | Benchmark dataset 1845-1901                              | 87        |
| 6.1.5.2  | Benchmark dataset 1787-1845                              | 88        |
| 6.2      | Rule-based linking in release 1                          | 89        |
| 6.2.1    | The algorithm  | 89        |
| 6.2.1.1  | Blocking   | 89        |
| 6.2.1.2  | Similarity scores  | 91        |
| 6.2.1.3  | Link score   | 91        |
| 6.2.1.4  | Applying thresholds (primary links)                      | 91        |
| 6.2.1.5  | Household links  | 91        |
| 6.2.1.6  | Iteration and tolerance adjustment                       | 91        |
| 6.2.2    | Link rates   | 91        |
| 6.2.3    | Quality testing  | 91        |
| 6.3      | Machine learning linking in release 2                    | 93        |
| 6.3.1    | Introduction to linking with supervised learning and XGB | 94        |
| 6.3.1.1  | Supervised learning                                      | 94        |
| 6.3.1.2  | Encoding   | 95        |
| 6.3.1.3  | Decision trees and XGB                                   | 96        |
| 6.3.1.4  | Post-calibration of XGB links                            | 98        |
| 6.3.2    | Overview of machine learning models used for linking     | 100       |
| 6.3.2.1  | Features   | 101       |

|          |   |            |
|----------|---|------------|
| 6.3.3    | Linking direction and blocking . . . . .  | 103        |
| 6.3.4    | Training and calibration data . . . . .   | 104        |
| 6.3.4.1  | Generation of additional non-links . . . . .  | 105        |
| 6.3.5    | Selecting threshold $\delta$ for post-calibration of XGB links . . . . .                                      | 107        |
| 6.3.5.1  | Census models C and C0 . . . . .  | 108        |
| 6.3.5.2  | Parish records and Copenhagen burials models . . . . .  | 108        |
| 6.3.5.3  | Resolving link conflicts due to multiple models . . . . .   | 108        |
| 6.3.6    | Results: Link rates . . . . .   | 109        |
| 6.3.6.1  | Overall link rates - per model per year range . . . . .   | 109        |
| 6.3.6.2  | Representativity assessment . . . . .   | 117        |
| 6.3.7    | Validation . . . . .  | 122        |
| 6.3.7.1  | Sampling . . . . .  | 123        |
| 6.3.7.2  | Validation guidelines . . . . .   | 124        |
| 6.3.7.3  | Results . . . . .   | 125        |
| 6.4      | Presentation of links in Link-Lives release 2 . . . . .   | 129        |
| <b>7</b> | <b>Life course aggregation</b>  | <b>132</b> |
| 7.1      | Self developed algorithm for rule-based record linkage in release 1<br>(also included in release 2) . . . . . | 132        |
| 7.2      | Life-course aggregation method for machine learning links in re-<br>lease 2 . . . . .                         | 132        |
| 7.2.1    | Selection of high quality models for aggregation . . . . .  | 132        |
| 7.2.1.1  | General approach to model selection . . . . .   | 133        |
| 7.2.1.2  | Specific models considered to perform well enough<br>to include in life-course aggregation . . . . .          | 133        |
| 7.2.1.3  | Results . . . . .   | 134        |
| 7.2.2    | Initial life-course aggregation . . . . .   | 135        |
| 7.2.3    | Life-course level sanity checks . . . . .   | 135        |
| 7.2.4    | Results . . . . .   | 137        |
| 7.2.5    | Life-course quality test . . . . .  | 137        |
| 7.3      | Presentation of the life-courses in Link-Lives release 2 . . . . .  | 140        |
| <b>8</b> | <b>Extracting data</b>  | <b>142</b> |
| 8.1      | How to combine harmonised and original transcribed datasets . . . . .   | 142        |
| 8.2      | How to combine links and LL harmonised datasets . . . . .   | 142        |
| 8.3      | How to combine life-courses and LL harmonised datasets . . . . .  | 143        |
| <b>9</b> | <b>Documentation on the construction of auxiliary datasets</b>  | <b>144</b> |
| 9.1      | ALA versions of data . . . . .  | 144        |
| 9.1.1    | Censuses . . . . .  | 144        |
| 9.1.2    | Parish registers . . . . .  | 144        |
| 9.1.3    | Copenhagen Burial Registers . . . . .   | 144        |
| 9.2      | Names . . . . .   | 144        |
| 9.2.1    | Synonym catalogue by Thomsen . . . . .  | 145        |
| 9.2.2    | Synonym catalogue by Revuelta-Eugercios and Kællerød . . . . .  | 145        |
| 9.2.3    | Methods for creating the Link-Lives name synonym catalogue . . . . .  | 147        |
| 9.2.4    | Description of the synonym catalogue included as part of<br>the release . . . . .                             | 148        |



|           |  |            |
|-----------|--|------------|
| 9.3       | Geography . . . . .  | 148        |
| 9.3.1     | Fitting the census geography to the existing administrative boundaries: the Danish census historical GIS . . . . . | 148        |
| <b>10</b> | <b>Benchmark dataset 1845-1901</b>   | <b>149</b> |
| 10.1      | Structure of the file . . . . .  | 149        |
| 10.2      | Content of the file . . . . .  | 149        |
| 10.3      | Data used for linking . . . . .  | 150        |
| 10.4      | Specific linking units included . . . . .  | 150        |
| 10.5      | Effects of linking interface on data creation . . . . .  | 151        |
| 10.5.1    | Overall changes in the software and documentation . . . . .  | 151        |
| 10.5.2    | Stability of the visualisation of variables . . . . .  | 152        |
| 10.5.3    | Availability of potential candidates . . . . .   | 153        |
| 10.6      | Summary of Best Practices . . . . .  | 154        |
| 10.7      | Definition of disagreements . . . . .  | 158        |
| 10.8      | Pipeline from links files to benchmark dataset . . . . .   | 167        |
| 10.9      | Training linkers . . . . .   | 169        |
| 10.9.1    | Selection and training of linkers . . . . .  | 169        |
| 10.9.2    | Linking school . . . . .   | 170        |
| 10.9.3    | First linking school design . . . . .  | 170        |
| 10.9.4    | Linking school sessions . . . . .  | 170        |
| 10.10     | Linking continuing monitoring: workshops and the “link-in” . . . . .   | 171        |
| <b>11</b> | <b>Benchmark dataset 1787-1840</b>   | <b>173</b> |
| <b>12</b> | <b>Codebooks</b>   | <b>174</b> |
| 12.1      | Codebook for censuses . . . . .  | 174        |
| 12.2      | Codebook for parish registers . . . . .  | 184        |
| 12.3      | Codebook for Copenhagen Burial Register . . . . .  | 201        |
| 12.4      | Codebook for LL harmonized data . . . . .  | 207        |
| 12.5      | Codebook for censuses for ALA . . . . .  | 212        |
| 12.6      | Codebook for parish registers for ALA . . . . .  | 215        |
| 12.7      | Codebook for Copenhagen Burial Register for ALA . . . . .  | 225        |
| 12.8      | Codebooks for links files . . . . .  | 229        |
| 12.9      | Codebook for life-courses . . . . .  | 231        |
| 12.10     | Codebook for synonym catalogues . . . . .  | 232        |
| 12.11     | Codebook for Benchmark dataset 1787-1901 . . . . .   | 233        |
| <b>13</b> | <b>Appendix</b>  | <b>236</b> |

## List of Figures

|    |  |     |
|----|--|-----|
| 1  | Overview of Link-Lives pipeline and relevant sections where they are described. . . . .  | 15  |
| 2  | An example of a household from the census of 1860, parish of Junget (Jutland) . . . . .  | 23  |
| 3  | An example of a household from a parish register in Kvanløse parish 1853-1854, births from girls . . . . .   | 32  |
| 4  | An example of a pre-printed and completed form from January 1896 in Copenhagen Burial Register, showing six individual burial records. . . . .   | 38  |
| 5  | The linking interface (ALA) with options to browse data, search the sources and make link-decisions. . . . .   | 84  |
| 6  | Link-rates of the Benchmark dataset 1787-1845 and 1845-1901. . .   | 88  |
| 7  | The rule-based algorithm. . . . .  | 90  |
| 8  | Total count of links from census to census and from Copenhagen burial register to census. . . . .  | 92  |
| 9  | Total count of links from PR to censuses. . . . .  | 92  |
| 10 | Link rates of links made from censuses and Copenhagen burial register. . . . .   | 93  |
| 11 | Example of decision three with input $v$ containing the number of different letters in <code>name</code> , the absolute distance of <code>birth_year</code> , and weather or not <code>event_district</code> is the same. Starting from the left, a number of decisions are made based on the features to produce the final prediction $\hat{z}$ . . . . . | 97  |
| 12 | Link probabilities for all record pairs from source A with $n = 3$ records and source B with $m = 4$ records. The numbers denote the link probabilities $p_{i,j}$ for each of the 12 possible pairs. . . . .   | 99  |
| 13 | Link Rates for Baptisms CFM model . . . . .  | 111 |
| 14 | Link Rates for CF model . . . . .  | 111 |
| 15 | Link Rates for CM model . . . . .  | 112 |
| 16 | Link Rates for Confirmations CFM model . . . . .   | 112 |
| 17 | Link Rates for Confirmations CF model . . . . .  | 113 |
| 18 | Link Rates for Confirmations CM model . . . . .  | 113 |
| 19 | Link Rates for Marriages GB model . . . . .  | 114 |
| 20 | Link Rates for Burials CFM model . . . . .   | 114 |
| 21 | Link Rates for Burials CF model . . . . .  | 115 |
| 22 | Link Rates for Burials CM model . . . . .  | 115 |
| 23 | Link Rates for Burials GB model . . . . .  | 116 |
| 24 | Link Rates for Burials 1P model . . . . .  | 116 |
| 25 | Link Rates for Burials in Copenhagen Burials 1P model . . . . .  | 117 |
| 26 | Ratio of female to male link rates for different census pairings . .   | 118 |
| 27 | Ratio of female to male link rates for different events, models and census years . . . . .   | 118 |
| 28 | Link rates by age for different census pairings . . . . .  | 119 |
| 29 | Link rates by age for confirmations for all target censuses (model CFM) . . . . .  | 120 |

|    |  |     |
|----|--|-----|
| 30 | Link rates by age for burials for target census 1801-1901 for all models)            | 121 |
| 31 | Link rates by age for burials in Copenhagen for target census 1860-1901 for model 1P | 122 |
| 32 | Examples of visualization of life-courses  | 135 |
| 33 | A life-course with two person appearances from the 1860 census                       | 138 |
| 34 | A life-course with two person appearances in births as main person                   | 139 |
| 35 | Size of life-courses   | 139 |
| 36 | Example of improvements in release 2   | 140 |
| 37 | Geographical distribution of linking units by origin source                          | 151 |
| 38 | ALA interface showing layout of primary and secondary variables                      | 153 |
| 39 | Decision tree aid for our domain experts.  | 158 |
| 40 | Pipeline from outputs of a linker to the benchmark dataset.                          | 168 |

## List of Tables

|    |   |     |
|----|---|-----|
| 1  | Original sources used in release 2, their year-range and creators of the transcriptions. . . . .  | 16  |
| 2  | Overview of datasets included in the downloadable Link-Lives release 2. . . . .   | 17  |
| 3  | Overview of datasets in Link-Lives release 2 that need to be requested. . . . .   | 18  |
| 4  | Danish census years and inclusion status . . . . .  | 25  |
| 5  | Main characteristics of the projects that created the transcriptions used by Link-Lives in release 2. . . . .   | 41  |
| 6  | Number of transcribed records in each census year (from versions delivered to Link-Lives). Note that census of 1885 was carried out only in Copenhagen. . . . . | 45  |
| 7  | Number of valid records included in Link-Lives version 2 from the Copenhagen Burial Register divided by decade. . . . .   | 56  |
| 8  | The standardised value of <b>civilstatus</b> in Danish and their respective translations in English. . . . .  | 59  |
| 9  | Source ids values used in Link-Lives. . . . .   | 62  |
| 10 | Number of harmonized records in censuses and Copenhagen burials   | 63  |
| 11 | Total number of individual person appearances in the parish registers (harmonized dataset), divided by decade and event type. . .                               | 64  |
| 12 | Roles used in the Link-Lives standardised data according to event and dataset. . . . .  | 66  |
| 13 | Codebook for sex in synonym catalogue (standard values). . . . .  | 73  |
| 14 | Values of marital status used in Danish and their translation to English. . . . .   | 74  |
| 15 | Main features and characteristics of methods in domain-expert linking of different sources. . . . .   | 78  |
| 16 | Overview of models . . . . .  | 101 |
| 17 | Models and source pairs. . . . .  | 101 |
| 18 | Model features overview . . . . .   | 102 |
| 19 | Specific support persons used by the models . . . . .   | 103 |
| 20 | Summary of blocking parameters and direction. . . . .   | 104 |
| 21 | Training Data Sources for Models. . . . .   | 105 |
| 22 | Selection of non-links for each link in labeled data. . . . .   | 106 |
| 23 | Blocking Approach for Different Models for Training Data Construction. . . . .  | 107 |
| 24 | Model C and C0 sample of links and non-links. . . . .   | 107 |
| 25 | Precision and Recall for Different Source Pairs Evaluated on Calibration Data. . . . .  | 108 |
| 26 | Precision, Recall, and Test Links from Parish Records to Censuses evaluated on Calibration data . . . . .   | 109 |
| 27 | Link Rates and Number of Linkable Records between Consecutive Censuses . . . . .  | 110 |
| 28 | Number of links validated in each sample . . . . .  | 124 |

|    |   |     |
|----|---|-----|
| 29 | Validation results for census→census links, with 95% confidence intervals . . . . .   | 125 |
| 30 | Example validation results by target census . . . . .   | 126 |
| 31 | Precision scores in validated samples for parish registers (correct links as % of all validated links in sample) . . . . .                                      | 127 |
| 32 | Description of method ids used in Link-Lives version 2. . . . .   | 130 |
| 33 | Number of links in release 1 and release 2 . . . . .  | 131 |
| 34 | Summary of omitted links by event type, model, and reason . . . . .   | 134 |
| 35 | Example of a life course in life-courses file. . . . .  | 141 |
| 36 | Transposed life-course no. 4. . . . .   | 143 |
| 37 | Number of person appearances with a decision by origin and target source. . . . .   | 151 |
| 38 | Coverage of origin sources by parish and target census year in the set of “core parishes” . . . . .   | 152 |
| 39 | Summary of the Best Practices main changes. . . . .   | 155 |
| 40 | Possible combinations of linker decisions and actions taken about them. . . . .   | 159 |
| 41 | Linking School Sessions held. . . . .   | 171 |
| 42 | Codebook for censuses. . . . .  | 175 |
| 43 | Codebook for Parish Registers. . . . .  | 185 |
| 44 | Codebook for Copenhagen Burial Register. . . . .  | 202 |
| 45 | Codebook for Link-Lives harmonized data. . . . .  | 208 |
| 46 | Codebook for ALA census data. . . . .   | 213 |
| 47 | Codebook for ALA parish register data. . . . .  | 216 |
| 48 | Codebook for ALA Copenhagen Burial Register data. . . . .   | 226 |
| 49 | Codebook for links from rule-based approach from release 1. . . . .   | 229 |
| 50 | Codebook for links from machine learning approach from release 2. . . . .   | 230 |
| 51 | Codebook for life-course files . . . . .  | 231 |
| 52 | Codebook for benchmark dataset 1787-1901. . . . .   | 234 |
| 53 | Overview of datasets included in the release . . . . .  | 236 |
| 54 | Overview of main datasets that can be requested . . . . .   | 237 |
| 55 | Number of person appearances in test dataset by origin linking unit (parish, street/neighbourhood or year) and target census for the period 1845-1901 . . . . . | 238 |
| 56 | Number of person appearances in test dataset by origin linking unit (parish, street/neighbourhood or year) and target census for the period 1787-1845 . . . . . | 242 |

# 1 Preface

Link-Lives is a cross-disciplinary research project that takes information relating to a given person, drawn from diverse historical sources, to create life-courses and family relations from 1787 to the present. It combines machine learning, historical research and citizen involvement to transform Danish archival sources into historical big linked data. It has been carried out in a partnership between the Danish National Archives (hereafter *Rigsarkivet*), the Copenhagen City Archives (hereafter *Københavns Stadsarkiv*) and the University of Copenhagen.

Link-Lives was funded for six years (2019-2025) through two grants, one from the Innovation Fund Denmark and one from the Carlsberg Foundation<sup>1</sup>. The data is made available in two ways:

- For researchers through a combination of free downloads of the majority of the datasets through *Rigsarkivet*'s data repository and request-only datasets for parts of the data subjected to legal protections.
- For the public through a searchable website in Danish, with limited download capabilities (restricted too only data that does not contain information on individuals protected by any personal information legislation): [link-lives.dk/soeg/](https://link-lives.dk/soeg/).

The project overall combines almost 75 million records from a variety of actors: two archives (*Rigsarkivet* and *Københavns Stadsarkiv*) provided the data produced in the context of crowdsourcing cultural heritage data projects; the company Ancestry made their data available free of use in Denmark for limited purposes; and the transcription of additional censuses, carried out by the project. The work has been carried out in two phases, divided chronologically between census data from before and after 1901, which entails different access conditions, given the need for considerations of data protection for the period after 1901.

Datasets in phase 1 are:

- Nationwide censuses completely transcribed at the start of the project (1787, 1801, 1834, 1840, 1845, 1850, 1860, 1880 and 1901 and 1885 covering just Copenhagen) in a volunteer transcription project hosted at *Rigsarkivet*.
- Nationwide parish registers, 1814-1917, transcribed by the company Ancestry
- Copenhagen Burial Register, 1861-1911, transcribed by a volunteer project hosted at *Københavns Stadsarkiv*.

Datasets in phase 2 are:

- Nationwide censuses from 1890, 1911, 1916 and 1921, transcribed within the Link-Lives project in cooperation with the company Rooftop, which implemented automatic transcription methods.

---

1. Grant number 8088-00034A and CF18-1116, respectively.

- Copenhagen Police Sheets (*Politiets registerblade*), 1890-1923, a volunteer-transcription project carried out at *Københavns Stadsarkiv*.

The aim of Link-Lives has been to harmonise and link these datasets together to create life-courses, using a variety of methods. The first data release along with the Link-Lives website (corresponding to phase 1 data) were published in 2022. Release 2, which is covered by this guide, consolidates data from Release 1, applies new linking methods and supplies additional documentation about the treatment of the data. While still at an advanced stage, the results of the project should still be considered a work in progress. A further release (Release 3) includes phase 2 data and links, and will complete the results of the project (expected early 2026).

The data and documentation produced by Link-Lives are the result of the combination of the talent, expertise and hard work of more than 50 people across the University of Copenhagen, *Rigsarkivet* and *Københavns Stadsarkiv*. Over the course of the project, historians and archivists have become more proficient in data science methodologies and programming and data scientists and data specialists have come to acquire historical competences and an appreciation for the complexity and richness of “messy” historical data. Together, we have bridged competences in a profile that we have named “Pystorians”, which is the portmanteau of “Python”, the project’s programming language, and “historians”, denoting that this is a project that places historical competences at the core of its development.

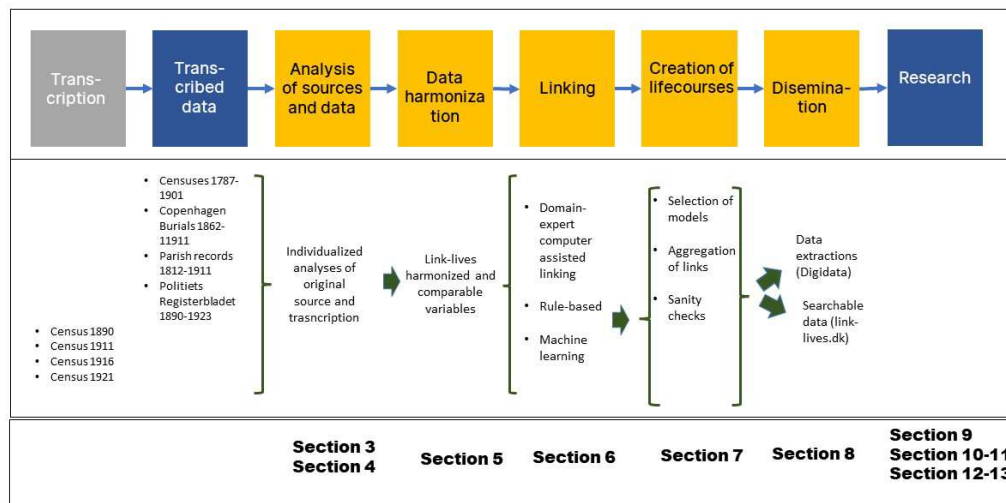
During this period, we have had advice and support of an advisory board (see full list of participants in the first page).

## 2 Introduction

### 2.1 Purpose and overview of the guide

This guide provides context for and methodological descriptions of the Link-Lives dataset release 2. We follow our own framework for describing data built from historical sources (Robinson et al. 2023, 152–172) and provide a brief description of all the stages of data transformation from the historical sources. Figure 1 shows the pipeline developed by Link-Lives following that framework and indicates the specific sections in this guide that refer to them. The sources displayed in the figure illustrate the universe of sources included in Link-Lives overall even though the additional censuses and textitPolitiets registerblade are not included in release 2. Each section has been mainly written by different members of the team with different core competences and the text reflects the many hands involved in it and their different styles. We do not anticipate anyone reading this guide from start to finish, but find it more likely to be used as a reference for all stages of the process, from sources to the concrete datasets we have produced. To aid look-up, we included references to the different sections through the guide, to lead readers to more details according to their needs.

Figure 1: Overview of Link-Lives pipeline and relevant sections where they are described.



The following sections introduce the historical sources in their context (section 3), describe the process of digitization that happened prior to their incorporation in Link-Lives (section 4), as well as the Link-Lives data processing (section 5) and the different approaches to linkage (section 6) and aggregating life-courses we have implemented (section 7). We also provide a section to help users combine the different datasets and create personalized extractions (section 8). Finally, there



are several sections with information about the additional auxiliary datasets that we have created in the previous sections (section 9) as well as comprehensive documentation on benchmark datasets for linking (sections 10 and 11), the result of the method described in section 6.1. The last part of the guide consists of codebooks for all the datasets included in the release (section 12). We have a final section with additional appendices (section 13).

In addition to this guide, we have made available as part of the Link-Lives release 2 additional documentation for the construction of the benchmark dataset that we call Link-Lives paradata. While metadata is data about the data, which we also made available in the files, the term paradata refers to the documentation and data pertaining to the method with which the data is created. We provide as paradata with several relevant working documents that were used to create domain-expert linked data ([Link-Lives Paradata](#)<sup>2</sup>).

## 2.2 Overview of release: datasets and availability

Release 2 contains life-courses linked across censuses (1787-1901), parish registers (1814-1917) and Copenhagen burial registers (1861-1911). The datasets have been transcribed by different methods, as can be seen in table 1. They have been linked with three different methods: domain-expertise, rule-based and machine learning.

Table 1: Original sources used in release 2, their year-range and creators of the transcriptions.

| Sources                    | Year range | Transcribed by  |
|----------------------------|------------|---|
| Census                     | 1787–1901  | Crowdsourcing project at <i>Rigsarkivet</i>           |
| Copenhagen burial register | 1861–1911  | Crowdsourcing project at <i>Københavns Stadsarkiv</i> |
| Parish registers           | 1814–1917  | Ancestry  |

## 2.3 Types of files and where to find them

We make available three types of data: source files, link life-course files, and auxiliary data files. The majority of the data is available through *Rigsarkivet's Digidata* direct download function but some files need to be requested.

In table 2 we provide a summary of the files downloadable from release 2 (see a full list in table 53 in the appendix in section 13) and in table 3 we provide

<sup>2</sup> The full name of the file is “Link-Lives Paradata. Computer Assisted Domain Expert Record Linkage.pdf”

a summary of the files that can be requested and from whom (a full list is also provided in the appendix in table 54).

Table 2: Overview of datasets included in the downloadable Link-Lives release 2.

| Folder                    | Content  | Files         | Description   |
|---------------------------|--|---------------|---|
| main_datasets             | Census (1787–1901) and Copenhagen burials datasets | 28 + 1 folder | Transcribed and harmonized versions.  |
| main_datasets/<br>ALA     | ALA-formatted datasets                             | 14            | Census + Copenhagen burials in ALA format.  |
| links_and_<br>lifecourses | Life courses and link datasets                     | 5             | Includes links and life-course files for release 1 and 2 and the benchmark dataset. |
| auxiliary                 | Synonym catalogues                                 | 3             | Name, sex, and marital-status synonym catalogues.                                   |

### 2.3.1 Source files

There are 17 sources: 10 censuses, six types of parish registers and one Copenhagen burial register and we provide three versions for each individual source.

- Original files: the transcribed data in almost exactly the form it was received by us with only added minimal information to ease connection with the rest of Link-Lives data (see 4). The files have a suffix `*_cl.csv`, which stands for “clean”, denoting that they are the clean datasets as delivered by data providers. See codebooks for each type of source in tables 42, 43 and 44.
  - The census data files are directly available from Digidata. They are extractions from *Dansk Demografisk Database* (Danish Demographic Database) from December 2019 (and early 2020 for 1901).
  - *Københavns Stadsarkiv* files include the original transcribed data as it was on March 2021 available for request from *Københavns Stadsarkiv* at [Arkivfinder](#). Data will be provided by e-mail after the applicant provides information on intended use.
  - Ancestry’s files include transcriptions of the Danish parish registers for the period 1814-1917 for all parish events (baptisms, marriages, confirmations, burials, arrivals and departures). They were extracted from Ancestry’s own systems in May 2021. Access to these are subject to a written request via *Rigsarkivet* at [data@rigsarkivet.dk](mailto:data@rigsarkivet.dk). Ancestry has given a general permission for research purposes for small scale extractions, which can be obtained within a few days. Full dataset access requires obtaining individual permission from Ancestry, which may incur a longer wait.

- Harmonized Link-Lives versions: these files have been created within the Link-Lives structure and only have standard values for a selection of the variables. They are characterized by ending in `*_std.csv` (standard structure) (see section 5).
  - Census data and *Københavns Stadsarkiv* files are directly available from Digidata.
  - Parish records files require a separate request submission to Ancestry, via *Rigsarkivet*.
- ALA files: these files are minimally processed and re-structured versions to be used with our software, ALA (Assisted Linking Application), for domain-expert linking. They have the prefix `ALA_`. See sections 9.1, 6.1 and 10 for further details. As they are so close to the original files, access is governed by the same restrictions as the original files.

Table 3: Overview of datasets in Link-Lives release 2 that need to be requested.

| Folder            | Content                                | Nr. Files | Description   |
|-------------------|--|-----------|---|
| main_datasets     | Copenhagen Burial Register (1861-1911) | 1         | Transcribed version can be requested from <i>Københavns Stadsarkiv</i> .  |
| main_datasets     | Parish registers (1814-1917)           | 10        | Transcribed and harmonized files. Data can be requested from <i>Rigsarkivet</i> , who handles the request to Ancestry.  |
| main_datasets/ALA | ALA-formatted datasets                 | 100+      | ALA formatted datasets for PR baptisms, marriages and burials, parish by parish. Data can be requested from <i>Rigsarkivet</i> , who handles the request to Ancestry. |

### 2.3.2 Link files

There are three types of files that connect individuals through different sources and they are all available through Digidata.

- The `links.csv` files: connections between person appearances from different sources produced by automatic methods. See sections 6.2 and 6.3 for methods, see codebooks in table 49 and 50.
- The `lifecourses.csv` files: constructed life-courses as a result of aggregating links. See section 7 for methods, table 51 for codebook.

- The `benchmark_v1.xlsx` file: includes all the linked person appearances created by our domain experts following our method. See section 6.1 and 10 for methods, see table 52 for codebook.

### 2.3.3 Auxiliary data

We have created a number of additional auxiliary datasets to support harmonisation and linking of datasets. We provide name, place and sex standardization synonym catalogues under folder `auxiliary`. They are prefixed by `SC`, which stands for “synonym catalogue”. A codebook of their structure is included in section 12.10. However, additional tools not included or extended versions can be requested by contacting *Rigsarkivet* at `data@rigsarkivet.dk`

### 2.3.4 Data formats

Link-Lives has worked almost exclusively with `*.csv` files and “,” as delimiters but for a few situations where the excel format has been used.

## 2.4 What's new in release 2?

We published Link-Lives release 1 in June 2022 on the `link-lives.dk` website and *Rigsarkivet's* website as a beta release. Release 2 features a new set of machine learning links, improved structure and documentation. It also has its own DOI. Moreover, it replaces release 1, which is no longer available for download. Release 2 is the first version of the Link-Lives data published in Digidata. It is still a work in progress, but has significant improvements to release 1.

- Release 2 uses exactly the same original data as in release 1 and it provides the same harmonisation.
- Overall, file structure and naming of datasets have changed to comply to Digidata requirements and make it more user-friendly (i.e., each census has a unique name instead of being within a set of folders).
- It includes an additional version of all the datasets, the ALA files, which can be used with our software ALA (see more in section 6.1.3.3).
- Aside from the release 1 links and life-courses, it includes a new set of machine learning links (and life-courses). See sections 6.3 and 7.
- It includes the manually linked person appearances we produced in release 1 plus an additional number of linked person appearances in their own file (see section 6.1 and 10), with all the accompanying metadata, so they can be used on their own.
- This guide includes detailed context information about the sources, datasets and the processes carried out in the project, as well as detailed codebooks for each of the file versions (see section 12).

- The release includes paradata documentation<sup>3</sup> under the **Dokumenter** folder. **Link-Lives Paradata.Computer Assisted Domain Expert Linkage.pdf** aggregates a variety of working documents that capture some of the more detailed methodological elements of our work in relation to the construction of the domain expert linked data.

## 2.5 A searchable overview of the release: linklives.dk

Link-Lives releases are also made available and searchable on the webpage linklives.dk. The webpage displays the same data as release 2<sup>4</sup> and can be used to see the relationship between the different files in action. Users can easily check how we represent links and life-courses from the release and it can be easily used to verify links and life-courses.

- Person appearances: Concatenating the **source\_id** and **pa\_id** with a slash in the form “12-11619813”, users can add the text into the link-lives URL to find the relevant representation of the person appearance. Users should type “https://link-lives.dk/soeg/pa/12-11619813”.
- Lifecourses: it follows the same procedure but users need to concatenate the number of release 2, it is “2.1” and the relevant **lifecourse\_id**, for instance “1806041”. Users should type “https://link-lives.dk/soeg/life-course/2.1-1806041” to access the full life-course.
- Family relationships and household co-residence can also be explored by clicking in the relevant buttons.

## 2.6 How to cite

**Guide:** Revuelta-Eugercios, B.A., Robinson, O., Mathiesen, N.R., Thomsen, A.R., Gronemann, S.T., Sunde, L.B, Nørskov, I.D., Sparre, A.L., Kallehauge, T., and Løkke, A. (2025) *Link-Lives Release 2 Guide*, Danish National Archives, Copenhagen City Archives and University of Copenhagen, Denmark.  
DOI: 10.5279/dk-ra-14001

**Data:** Link-Lives. (2025) *Link-Lives release 2*, Danish National Archives, Copenhagen City Archives and University of Copenhagen, Denmark.  
DOI: 10.5279/dk-ra-14001

---

3. Paradata is an increasingly used concept to describe the process of creating data. While metadata is data about the data, paradata covers the process of data creations. There are multiple definitions of paradata but a very broad one can be found in latest monography about it *Paradata. Documenting Data Creation, Curation and Use* (2025), where Sköld (2025) writes that paradata “is information describing the practices and processes involved in how forms of information are created, curated and used.”

4. The only difference is that the webpage does not show all the links from the omitted models for life-course aggregation described in section 7.2.1.

## 2.7 License

All rights to the Link-Lives dataset are owned by the partners and contributors to Link-Lives.

### 2.7.1 All sources

- You may download data for personal use, for educational purposes and for use in research.
- You may not use data in any commercial activity.

### 2.7.2 The fully transcribed censuses 1787-1901

These datasets are already freely available for research and educational purposes and personal use through DDD (*Dansk Demografisk Database* at *Rigsarkivet*) and are available for free download via *Rigsarkivet's* website "[Folk i Fortiden](#)".

### 2.7.3 Transcription of parish registers

Ancestry has made available the transcriptions for the use of *Rigsarkivet's* and Link-Lives' users on condition they are not downloaded in their entirety and that Ancestry and *Rigsarkivet* are credited when they are used (e.g. in publications). It is necessary to sign a specific license with Ancestry to access larger data extractions. The National Archives handles those requests (see section [2.3](#)).

### 2.7.4 The Copenhagen Burial Register 1861-1911

This data is freely available for research and educational purposes as well as for personal use. The data may not be used commercially and may not be transferred to a third party or published as a whole. *Københavns Stadsarkiv* must be credited appropriately (see section [2.3](#)). Newer extractions of data can be applied for directly from *Københavns Stadsarkiv*.

### 2.7.5 Links, life-courses and the Link-Lives harmonized versions of the above data

These can be freely downloaded and used as long as the use respects the conditions for the above data. Link-Lives must be credited when the data is used (e.g. in publications). The release number of the Link-Lives data used must always be a part of the citation (see section [2.6](#)) and a copy of the publication must be sent to [data@rigsarkivet.dk](mailto:data@rigsarkivet.dk).

## 2.8 Conventions used in this guide

For ease of understanding we use the following conventions:

- **Bold** to denote a variable name, in contraposition to discussing as an attribute. For instance, the **age** variable does not include precise information about age

- `Monospace` font to denote file names part of the release
- *Italics* for terms in Danish
- “Quotes” for either emphasis or describing specific values of variables.

## 2.9 Contact

The project ended on 31 August 2025 but personnel and work at (Rigsarkivet) continues in similar projects. Contact in relation to access or minor questions about the data is possible by writing to [data@rigsarkivet.dk](mailto:data@rigsarkivet.dk) but there are no fixed resources allocated to this task, so please be aware that reply time may vary.





individual persons was not produced or preserved in any systematic way. In 1787 this was rectified, and from then on Danish censuses were conducted using standard forms, which recorded names and other individual level information. These original returns, handwritten on printed forms, are generally very well preserved at *Rigsarkivet*. See further description in *Hovedresultaterne af Folketællingen i Kongeriget Danmark den 1ste Februar 1890, med tilhørende Befolkningskaart* (1894), and Degn (1991, 7).

The last manually-recorded census was taken in 1970, two years after the Danish Civil Registration System (*Det Centrale Personregister, CPR*) was established in 1968. CPR-based register counts were taken in 1976 and 1981. Table 4 lists all censuses carried out in the period and shows their coverage (although censuses were nationwide, there were some additional Copenhagen censuses), transcription status (in bold those transcribed) and whether or not they form part of Link-Lives. Original materials have been preserved for all but 1935, 1945 and 1955, which were discarded. (Rosen 1991, 1830).

Table 4: Danish census years and inclusion status

| Year        | Coverage        | Link-Lives v2 | Link-Lives v3 |
|-------------|-----------------|---------------|---------------|
| <b>1787</b> | Nationwide      | Yes           | Yes           |
| <b>1801</b> | Nationwide      | Yes           | Yes           |
| <b>1834</b> | Nationwide      | Yes           | Yes           |
| <b>1840</b> | Nationwide      | Yes           | Yes           |
| <b>1845</b> | Nationwide      | Yes           | Yes           |
| <b>1850</b> | Nationwide      | Yes           | Yes           |
| 1855        | Nationwide      | No            | No            |
| <b>1860</b> | Nationwide      | Yes           | Yes           |
| 1870        | Nationwide      | No            | No            |
| <b>1880</b> | Nationwide      | Yes           | Yes           |
| <b>1885</b> | Only Copenhagen | Yes           | Yes           |
| <b>1890</b> | Nationwide      | No            | Yes           |
| 1895        | Only Copenhagen | No            | No            |
| <b>1901</b> | Nationwide      | Yes           | Yes           |
| <b>1906</b> | Nationwide      | No            | No            |
| <b>1911</b> | Nationwide      | No            | Yes           |
| <b>1916</b> | Nationwide      | No            | Yes           |
| <b>1921</b> | Nationwide      | No            | Yes           |
| 1925        | Nationwide      | No            | No            |
| 1930        | Nationwide      | No            | No            |
| (1935)      | Nationwide      | No            | No            |
| 1940        | Nationwide      | No            | No            |
| (1945)      | Nationwide      | No            | No            |
| 1950        | Nationwide      | No            | No            |
| (1955)      | Nationwide      | No            | No            |
| 1960        | Nationwide      | No            | No            |
| 1965        | Nationwide      | No            | No            |
| 1970        | Nationwide      | No            | No            |

Note: Census years in bold had been fully transcribed at the time of writing or transcribed as part of Link-Lives (1890, 1911, 1916 and 1921). 1906 and 1870 were completed after Link-Lives started. Years in parenthesis have been discarded.

The 1787 census was held on 1 July. All censuses from 1801 to 1925 were held in February, mostly February 1st. Winter was chosen because more people were at home with their family during winter (Degn [1991](#), 7). In summer children were hired to herd cows, sheep and geese, seafaring people were off shore, civil servants and construction workers were often traveling for business and well off people for leisure.

### 3.1.3 Who was registered where

From 1787, a census was to include each and every person (*alle og enhver*), natives as well as foreigners, who on census day were present in the kingdom, as well as all natives who were traveling abroad, but expected to return. Any person who had slept in a given dwelling the night before the census day was to be registered as belonging to that household or as temporarily visiting. Any person belonging to a given household, but who had not slept there that night, was to be registered as temporarily absent. Supplementary printed lists (*tillægslister*) were employed to handle the temporarily visiting persons and the temporarily absent, but over time it differed, which of these two groups should be recorded on the main lists and on the supplementary lists. See details in (*Hovedresultaterne af Folketællingen i Kongeriget Danmark den 1ste Februar 1890, med tilhørende Befolkningskaart* [1894](#), I–XL).

### 3.1.4 Geographical coverage

The geographical area, where censuses were held, changed over time, along with shifting geopolitical landscapes, and also differed from census to census. E.g. the 1787 census was only held in what was then called the kingdom of Denmark, excluding most parts of the composite state of the Danish king, such as the kingdom of Norway, Iceland and the Faroe Islands. These were, however, included in the 1801 census. The Duchies of Schleswig and Holstein had their own census in 1803. The 1834 census was the first that included Greenland. In 1835, censuses were held in the Danish West Indies, Tranquebar and Frederiksnagore (Serampore) in India, and in the Danish possessions on the coast of Guinea. A comprehensive examination of when and where in the composite state censuses were held can be found in *Hovedresultaterne af Folketællingen i Kongeriget Danmark den 1ste Februar 1890, med tilhørende Befolkningskaart* ([1894](#), I–XL).

The area of modern Denmark called *Sønderjylland*, on the border with Germany, has a complicated geopolitical history, which is reflected both in the censuses taken and in the preservation status of the original census lists. Until 1864 *Sønderjylland* was part of the Duchy of Schleswig, which was part of the composite state of the Danish king. After Denmark lost the war in 1864, almost all of *Sønderjylland* together with the rest of Schleswig came under Prussian rule until 1920, when the population at a referendum vote decided to become a part of Denmark. Prior to 1920, therefore, censuses for *Sønderjylland* are different from the ones taken in the kingdom of Denmark and many of the original handwritten lists did not survive archival selection or acts of war. A brief description about which lists remain and where they are kept is given in Degn ([1991](#), 18) and more thoroughly in Hertz ([1970](#)). Information about the censuses held at *Rigsarkivet*

can be found in Rosen (1983, 360) and Rosen (1991, 1942). Link-Lives release 2 includes those *Sønderjylland* census returns that were transcribed as part of the Danish Demographic Database.<sup>5</sup>

### 3.1.5 Legal basis and instructions for the registration

The process for carrying out a census was established by law, followed by written instructions to the civil servants charged with enumeration. The census of 1787 was decreed by the king in a special law (*Kongelig Resolution af 21. Marts 1787*) followed by *Reskript af 11. Maj 1787* sent to *Kjøbenhavns Magistrat* (Copenhagen City council), til *Stiftamtmandene* (the district magistrates) and the bishops. The *rescript* included a plan for the organization of the census, design of the printed forms and instructions for how to complete them. Such individual laws were also given for the next censuses. In 1835 a more general law (*Resolution af 9. juni 1835*) proscribed that censuses should be taken every fifth year, so from then on only the *rescript* (later called a *cirkulære*) was required to trigger the process. These instructions are important for modern users of the recorded censuses but they are difficult to find. We have found neither a total overview of archival/bibliographical references nor a total collection of archived laws, *reskripter* or *cirkulærer*. The most comprehensive overview of the content of laws and instructions 1769 to 1890 is in *Hovedresultaterne af Folketællingen i Kongeriget Danmark den 1ste Februar 1890, med tilhørende Befolkningskaart* (1894, I–XL), but even there there are no references to all the full text laws and instructions. Archival references to some of the instructions have been found by Nordholt Aagaard (2024) and more are expected to be published in Hjorth-Moritzsen (2026) .

### 3.1.6 The layout of the pre-printed forms and information recorded

The information recorded in each census and the layout of the printed forms varies slightly from year to year, but the core variables remain remarkable stable throughout the nineteenth century. The changes over time mostly concern the summary front pages and the supplementary forms (*tillægslister*) to record temporary visitors and temporarily absent people (*Hovedresultaterne af Folketællingen i Kongeriget Danmark den 1ste Februar 1890, med tilhørende Befolkningskaart* 1894, I–XL). Metadata, such as the census date, name of town and street, parish, district, county (*sogn, herred, amt*) also moved around the pages over time. In the early censuses this information was often found in page headers; later more often on front pages.

The printed forms for each census have three slightly different lay-outs adapted for use in Copenhagen, provincial towns and rural districts, respectively.

The information about the individual persons is listed, household by household, in the main census form (*hovedliste*). Each person was recorded on a single row, normally in a specific order. For a typical household this was: head of household, spouse, their children (oldest first), other relatives then domestic ser-

---

5. A status of what has been transcribed can be found at a [volunteer site](#) associated with the project.

vants. In households of master artisans and other larger households, craftsmen, other higher ranking employees, pupils and apprentices were listed in between the family and the domestic servants. Throughout the period, the forms always included place, family number, name, age, civil status, position in household and occupation. More columns were added over time. The below descriptions are based on the images in *Arkivalieronline* and the experiences of the Link-Lives team working with the information in the transcribed censuses:

- Place: The first column in the census forms asked for supplemental information to the metadata given about the location where the form was completed. The type of information registered in this column is therefore very varied. In Copenhagen, for example, we find apartment numbers, floors and locations within buildings; in towns we may find the name of the house or type of building (e.g. mill). In rural areas, the names of hamlets, manors, institutions or farms were often written.
- Family number: The aim of this column was to capture information about the total number of households in the country and to give precise information about which people belonged to which household. However, the way enumerators completed this column varied significantly. For example, some left the column empty, some gave the same number to all the individuals in a household, some restarted a number series for each new household and others used the column to write a sequential id for the whole census sheet. Sometimes the column is even used to write “N” or “-”, “+” for individuals temporarily absent or present in the household.
- Name: Recorded in one column, mostly a first name followed by middle and surnames, but in some cases the surname was listed first. Married women have either their married surname or their maiden name recorded, or both e.g. *Marie Hansen født Jensen* where *født* means “born”.
- Age: Given in years or for infants sometimes in months, weeks or days. From 1787 to 1860 an age was rounded up to the subsequent birthday. From 1870, the age at the previous birthday was recorded. From 1901, **date of birth** replaced age.
- Civil status: This was mostly recorded using abbreviations for the main categories: U = *ugift* (unmarried), G = *gift* (married), E = *enke/mand* (widow/er). The less common categories were often fully written out, e.g. *separeret* (separated), *skilt/fraskilt* (divorced) and *forladt* (abandoned).
- Position in household and occupation: In 1787 and 1801, the form had a column for each of these two variables. However, from 1834 to 1860, and again in 1880, the forms recorded both categories in one and the same column, though often only one of the two was actually recorded. In some cases this was because the enumerator felt that one word covered both. E.g. *gaardmand* (farmer) listed first in a household mostly also indicates “head of household”. For “wife” and “child” it was often the other way around, that the enumerator might sometimes have seen their position in

the household as their occupation as well (farmer's assistant). In 1870 and from 1885, **position in household** and **occupation** were again recorded in separate columns.

- Place of birth: This column was introduced in 1845. Instructions in the header required, for the rural districts, that the sequence *sogn*, *amt* (parish, county) be written, for towns the name of the town and for foreign countries just the name of the country. However, sometimes the place of birth was given in less precise ways such as just the county or a large region of the country such as *Jylland* (Jutland). The expression *her i sognet* (here in this parish) was often used when a person was born in the parish where the census was being taken.
- Køn (gender/sex): The registration of this variable started in 1870. It was to be filled with the initials M = *mand* (male) and K = *kvinde* (female). These two single letters, in contemporary handwriting, can sometimes be difficult to distinguish even for experienced readers.
- Religious community: This first appeared as a column in 1855 and the last appearance was in 1921. An overwhelming majority of the population (99.3 per cent in 1855 and 98.5 per cent in 1890) belonged to the Danish *Folkekirke* (*Befolkningsforholdene i Danmark i det 19. Aarhundrede. Statistisk Tabelværk* 1905, 71), which is the established Lutheran church, governed by the king until 1849, and since then by parliament and constitutional government. *Folkekirken* had (and has) its own income, regulated by law. However, there were multiple ways of spelling this adscription so there is substantial variety in spellings and denominations even within this relatively homogeneous religious population.

### 3.1.7 Other variables

In addition to the above variables, many censuses contained other columns. Some of these variables only appeared in one or two censuses and most of these have not been transcribed. Though it is possible to see them in the images of the originals held at *Arkivalieronline* or to look into the published, aggregate census statistics (see below).

Some of these variables are described below:

- Disabilities. There were ongoing questions relating to this over time. E.g. Is this person bereft of reason (*forstanden berøvet*)? If so, since the birth or at a later point in time? Also blindness and deafness were recorded, and whether since birth or at a later point in time.
- Net worth (*formue*) and income
- Year of wedding for the latest marriage
- Where this person lived precisely one year ago
- The number of alive children and the number of deceased children within this marriage union.

- The number of rooms with windows in the apartment/house
- Rent paid for housing for the latest half year.

### 3.1.8 Published, aggregate census results and analyses

The handwritten census returns were processed and published by contemporary statisticians, beginning with the 1834 census. The publication series *Statistisk Tabelværk* holds tables with all the data from the Danish censuses, aggregated into a range of distributions. A first chapter presents the set up of the actual census and gives analyses and discussions of the census results, often compared with earlier censuses. The data is well structured, comprehensive, informative and the statistical methods used are robust, even by today's standards.

As these publications provide easy access to high quality aggregate census data, they are very useful for users of the Link-Lives census data. E.g. it is possible to compare the number of inhabitants of a provincial town in a Link-Lives data extraction with the number of inhabitants given in the published census results.

Statistics Denmark (*Danmarks Statistik*) has scanned and published the whole series of volumes holding aggregate census results online: [Statistisk Tabelværk](#).

The aggregate statistics for the 1787 and 1801 censuses were not published at the time, but crude results are also available for them on the *Danmarks Statistik* website. The 1801 report can be found in the same volume as 1834 and 1787 as a scan of the original handwritten manuscript.

Also very useful is the demographic analysis of Denmark throughout the nineteenth century by Statistics Denmark, published in 1905 (*Befolkningsforholdene*). It is based on all the censuses as well as on all the published vital statistics in a revised and discussed form.

### 3.1.9 Further reading

- All Danish census results, in their the official, aggregate form published in the series *Statistisk Tabelværk*. Accessible via [Statistics Denmark's website](#).
- *Befolkningsforholdene i Danmark i det 19. Aarhundrede. Statistisk Tabelværk* 5. rk., Litra A, nr. 5. København: Statens Statistiske Bureau, 1905. Accessible via [Statistics Denmark's website](#).
- *Folketal, areal og klima 1901-60. Statistiske Undersøgelser* Nr.10, København, Det Statistiske Departement, 1964. Accessible via [Statistics Denmark's website](#).
- *Dansk kulturhistorisk opslagsværk*, Alstrup, Erik, and Poul Erik Olsen (ed.), "Folketællinger" 244-46, København: Dansk Historisk Fællesforening, 1991.
- Degn, Ole. *Alle skrives i mandtal: folketællinger og deres brug*. Arkivernes informationsserie. Kbh: Rigsarkivet, 1991. Accessible via [Rigsarkivet's website](#)

- “Fremgangsmaade ved Folketællingerne i Danmark i Tidsrummet 1769-1890” in *Hovedresultaterne af Folketællingen i Kongeriget Danmark den 1ste Februar 1890 : med tilhørende Befolkningskaart*. (Statistisk Tabelværk, Rk.4. Litra A ; Nr. 8, a) I-XLI, København, Det Statistiske Bureau, 1894. Accessible at [Statistics Denmark](#)
- Hertz, Michael. *De slesvigske og holstenske folketællinger 1803-1860*. Arkiv III (1970-71): 49-61.
- Holck, Axel. *Dansk Statistiks Historie 1800-1850 saerlig med Hensyn til den officielle Statistiks Udvikling*. København: Statens Statistiske Bureau, 1901. Accessible at [Statistics Denmark website](#).
- Salmonsens Konversations Leksikon vol VIII, (Chr. Blangstrup ed.), “*Folketælling*”, 387-90. København: Schultz Forlagsboghandel, 1919.
- Johansen, Hans Chr. *Danish population history 1600-1939*. University Press of Southern Denmark studies in history and social sciences; vol. 254. Odense: University Press of Southern Denmark, 2002.
- Marker, Hans Jørgen. *Danmarks befolkning 1801: Analyse på grundlag af folketællingen som mikrodata*. University of Southern Denmark studies in history and social sciences, vol. 507. Odense: Syddansk Universitetsforlag, 2015.
- Rosen, Wilhelm von (ed.). *Rigsarkivet og hjælpemidlerne til dets benyttelse I*. 2 vols. København: Rigsarkivet, 1983 360-364
- Rosen, Wilhelm von (ed.). *Rigsarkivet og hjælpemidlerne til dets benyttelse II 1848-1990*. 4 vols. København: Rigsarkivet, 1991 1826-1835

## 3.2 Parish Registers

The Danish parish registers contain the official registration of births, deaths and marriages as well as the religious events of baptisms, confirmations, weddings and burials. The official name for parish registers in Danish is *ministerialbøger*, but in daily life they are always called *kirkebøger*, which literally translates to “church books”. From 1814, it was mandatory for all parishes to keep two identical *ministerialbøger* in separate locations to prevent destruction by fire or flood. The main parish register was called *hovedministerialbog* and the copy’s name was *kontraministerialbog*; both consisted of pre-printed forms to be filled in by the parish priests and his aide. Link-Lives release 2 data is derived from the *kontraministerialbøger* 1814 to 1917 <sup>6</sup>.

### 3.2.1 Availability of images

At least one of the copies of nearly all Danish parish registers from 1814 has survived. They are kept at the Danish National Archives, and are available

---

6. See section [4.2](#) for more details on coverage



Figure 3: An example of a household from a parish register in Kvanløse parish 1853-1854, births from girls

| No. | Aar og Dato. | Barnets fulde Navn. | Dødsdato eller dødsårsag. | Børstet Navn, Dødsdato eller dødsårsag. | Børstet Navn, Dødsdato eller dødsårsag. | Børstet Navn, Dødsdato eller dødsårsag. | Børstet Navn, Dødsdato eller dødsårsag. | Børstet Navn, Dødsdato eller dødsårsag. |
|-----|--------------|---------------------|---------------------------|---|---|---|---|---|
| 8   | 1853         | Over Marie          | 1853                      | Over Marie                              | 1853                                    | Over Marie                              | 1853                                    | Over Marie                              |
| 9   | 1853         | Over Marie          | 1853                      | Over Marie                              | 1853                                    | Over Marie                              | 1853                                    | Over Marie                              |
| 10  | 1853         | Over Marie          | 1853                      | Over Marie                              | 1853                                    | Over Marie                              | 1853                                    | Over Marie                              |
| 11  | 1854         | Over Marie          | 1854                      | Over Marie                              | 1854                                    | Over Marie                              | 1854                                    | Over Marie                              |
| 12  | 1854         | Over Marie          | 1854                      | Over Marie                              | 1854                                    | Over Marie                              | 1854                                    | Over Marie                              |

The columns contain information on the children's birth dates, names, baptism dates, parents' names and residence, and godparents' names and residence.

online as black and white images scanned from microfilm (see [Arkivalieronline](https://arkivalieronline.dk) at [The Danish National Archives](https://arkivalieronline.dk)).

Newer, high-quality colour photographs of the *kontraminsterialbøger* 1814-1892 have been made by Ancestry and these images can be also be viewed on the Danish National Archives' website. The series is also freely accessible via Ancestry's Swedish website ([Ancestry.se](https://www.ancestry.se)), where the images appear side by side with Ancestry's transcriptions.

### 3.2.2 History

From 1646, it was mandatory for all parish priests to keep parish registers. After the Lutheran reformation, all Danish parishes had a Lutheran parish priest allocated by the Danish king, the head of the Danish Lutheran church. In the seventeenth and eighteenth centuries more than 99 per cent of the population belonged to this church even as a few other small religious communities existed. Until 1814 only one copy of the parish register was mandatory, and every priest was free to design it as he pleased. These older parish registers vary greatly, therefore, and some have been lost. (Ørberg 1991, 4-9; Johansen 2002, 14-18)

The *Reskript* 11. december 1812 decreed that from 1814 all parish registers

were to exist in duplicate using pre-printed, standard forms, so from this date, all Danish parish registers are fairly standardised (Ørberg 1991, 11; “Kirkebøger” 1991, 240–242)

During the nineteenth century, the vast majority of the Danish population still belonged to the Danish Lutheran church (now called *Folkekirken*). A small part of the population (1850: 0.5%, 1901: 1.4%) belonged to other officially-recognised religious communities, such as Jewish, Roman Catholic or Methodist (*Befolkningsforholdene i Danmark i det 19. Aarhundrede. Statistisk Tabelværk* 1905, 71). These communities kept their own community registers (“Kirkebøgerne” 1922, 931).

*Bekendtgørelse om Ministerialbøgers førelse 31. oktober 1891* and the *Instrux* of same date “Kirkebøgerne” (1922, 931)) introduced new printed forms from 1892, which required more detailed personal information and more thorough documentation of the identity of the registered persons. E.g. for marriages, the forms before 1892 only required the names, ages, occupations and places of residence for groom and bride but the new forms also require that the parish priests verify the full names, birth dates and birth places by inspecting birth or baptism certificates, where possible. The names of parents of both the bride and the groom and information about any previous spouses were also to be registered. For births, the parish priests were asked to verify the identity of the parents by inspecting their birth or baptism certificates and the legitimacy of the birth by marriage certificate, or by cross-referencing the parish records.

These changes mean that, after 1892, the parish registers are considered to be very well suited to identify individuals as they contain full name, birth date and birth parish of nearly all registered persons.

### 3.2.3 Geographical coverage

Rigsarkivet holds parish registers 1814 to 1918 covering the entire territory of today’s Denmark, including the Faroe Islands, Greenland and the southern part of Jutland also for the years 1864–1920, while under German rule. The registration practices of the latter, for the period 1875–1920, were different from the rest of the country, however. The legal registration of births, marriages and deaths was recorded in a separate civil register, while the aim of the *kirkebøger* primarily were to register the religious events of baptisms, confirmations, weddings and burials. However, information about birth and death were also included in the *kirkebøger* (“Kirkebøger” 1991, 442). Images of both registers are shown in Arkivalieronline (<https://arkivalieronline.rigsarkivet.dk/da/collection/theme/1>).

### 3.2.4 Information recorded

The parish registers are bound volumes divided into chapters by the type of registered events: births (*fødte*), confirmations (*konfirmerede*), marriages (*viede*) and deaths (*døde*). There are additionally records of arrivals (*tilgang*) and departures (*afgang*) in and out of parishes for the years 1814–1875 and for the same years a chapter called “*jævnførelseslister*”, meant to be a person index, but very often not kept properly (Dansk Kulturhistorisk Opslagsværk p 441).

Each parish register chapter has its own layout of the pre-printed form. Pages and entries are consecutively numbered. The below descriptions of the column headings are derived from inspection of a selection of pre-printed forms in Arkivaliononline before and after 1892. Some parish priests sometimes wrote much more information than the column headings asked for (e.g. cause of death). Not all variables in the originals have been transcribed by Ancestry.

**Births** Births are divided into two sub-chapters by headings on the pre-printed forms indicating whether records contain males or females. The headings reads “*Fødte Mandkøn*” (born males) and “*Fødte Kvindekøn*” (born females)

Each birth entry includes information on the child and its parents:

- Date of birth and date of baptism.
- Full name of the child, including surname. However, prior to 1828, a significant number of records only held the children’s given names.
- Names, residence and occupation of the parents in one single column. E.g. ”carpenter Jens Hansen and wife Kathrine Madsen, Torup (parish of residence)”.
- Names, residence and occupation of the godparents, similarly in one column.
- Mother’s age sometimes was recorded at the end of the column where the information about the parents was recorded. This started to happen around the mid-19th century.
- Parents’ birth dates and the date and place of their marriage began appearing in the parental information column from 1892.

**Confirmation** Confirmations are similarly divided by sex in the registers and are found in the chapters “*Confirmerede Dreng*” (confirmed boys) and “*Confirmerede Piger*” (confirmed girls).

Each confirmation record contains:

- Confirmation date.
- Name, age and residence of the confirmand.
- Names, civil status, residence and occupations of the parents, all in one column. This column could also contain information about step- or foster parents or the head of the household to which the confirmand belonged.
- Parish priest’s evaluation of the confirmand’s behaviour and religious knowledge.
- Date of vaccination and name of the vaccinator (disappeared during the 1870s after a new law of vaccination was introduced in 1871).
- Date and place of birth of the confirmand, mandatory from 1892.

**Marriages** Marriages are found in the chapters either entitled ”*copulerede*” earlier in the nineteenth century or ”*ægteviiede*” from 1892 onward.

Each marriage record contains:

- Marriage date and if the wedding took place in the church or at home. If at home the date of the permission given.
- Names, ages, civil status, occupations and place of residence of the bride and groom, with a column for the information of each of them.
- Name, occupation and residence of the two best men (*forlovere*).
- Date and place of birth for both the bride and groom appeared from 1892.
- Names and occupations of their parents were also recorded after 1892 as well as details of names of any previous spouses, and date of the previous spouse’s death or date of the divorce and permission to remarry.

**Deaths** Deaths are divided by sex and can be found in the chapters ”*Døde Mandkøn*” (deceased males) and ”*Døde Kvindekøn*” (deceased females).

Each death record contains:

- Date of death, and after 1892 also the place of death.
- Date of burial, and after 1892 also the place of burial.
- First and last name(s) of the deceased.
- Civil status of the deceased.
- Occupation and place of residence of the deceased.
- Age of the deceased.
- Sometimes before 1892 also the names of the father, if a child was legitimate, and the mother if illegitimate. In some cases also the name of the spouse, primarily for married women.
- After 1892, information on full name as well as place of birth and the names of the parents and spouse were required, if the information could be procured. Some parish priest also wrote the date of birth.

**Departures and arrivals** Departures and arrivals in the parish are found in the chapters ”*Afgangsliste. Afgaaede fra Sognet formedelst Bortreiste*” (departures) and ”*Tilgangsliste. Ankomne til Sognet*” (arrivals). The information in these chapters is of varying content, quality and coverage and has until now been difficult to use for research. However, they may prove to be much more usable in their searchable transcribed form.

Departures and arrivals records often contain:

- Date of the departure/arrival
- Name and age of the migrant
- Name of the parish they are moving to/from.
- The occupation of the migrant is often mentioned.

### 3.2.5 Published, aggregate vital statistics

From 1835, Danish vital statistics were published every five years in the series *Statistisk Tabelværk*. These volumes are based on the parish registers. The parish priests were obliged to send yearly counts of the vital events births, marriages and deaths to the bishops, who summarized them and forwarded the material to the central administration (*Rentekammeret*, later *Statistisk Bureau*). These handwritten lists were processed and published by contemporary statisticians aggregated in a range of distributions. A first chapter presents, analyses and discusses the data. The data is well structured, comprehensive, informative and the statistical methods used are robust, even by today's standards.

As these publications provide easy access to high quality aggregate vital statistics, they are very useful for users of the Link-Lives parish register data. For example, it is possible to compare the number of births, marriages and deaths in a main part of Denmark in a Link-Lives data extraction, with the number given in the published vital statistics.

The original printed volumes are scanned and published online by Statistics Denmark (*Danmarks Statistik*). The volumes for the years 1850-1889 have the titles *Vielser, Fødsler og Dødsfald i Aarene 18\*\*-18\*\**. The volumes for the years 1890-1969 are on a separate webpage under the title *Ægteskaber, fødte og døde*. The vital statistics for 1835 to 1849 are published together with the *censuses for 1834, 1840 and 1845*.

Also very useful is the analysis of the Danish demography throughout the nineteenth century done by Statistics Denmark and published in 1905 (*Befolkningsforholdene i Danmark i det 19. Aarhundrede. Statistisk Tabelværk 1905*). It is based on all the censuses as well as on all the published vital statistics in a revised and discussed form.

Similarly valuable for comparison are the original handwritten yearly reports to the central authorities made by the parish priests with counts of births, marriages and deaths. They are fully preserved for the years 1835-1868 at *Rigsarkivet* (arkivskaber *Rentekammeret* 1835-47 and *Statistisk Bureau* 1848-68, 1887)(Løkke 1998, 33-34, 137).

### 3.2.6 Further reading

- *Befolkningsforholdene i Danmark i det 19. Aarhundrede. Statistisk Tabelværk* 5. rk., Litra A, nr. 5. København: Danmarks Statistik 1905. <https://www.dst.dk/da/Statistik/nyheder-analyser-publ/Publikationer/VisPub?cid=19256>
- *Befolkningsudvikling. Befolkningsudvikling og sundhedsforhold 1901-1960*. Statistiske Undersøgelser Vol. 19. København: Det Statistiske Departement, 1966. <https://www.dst.dk/da/Statistik/nyheder-analyser-publ/Publikationer/VisPub?cid=19335>
- Holck, Axel. *Dansk Statistiks Historie 1800-1850 saerlig med Hensyn til den officielle Statistiks Udvikling*. Copenhagen: Statens Statistiske Bureau, 1901.

- Johansen, Hans Chr. *Danish population history 1600-1939*. University Press of Southern Denmark studies in history and social sciences ; vol. 254. Odense: University Press of Southern Denmark, 2002.
- "Kirkebøger" In *Dansk kulturhistorisk opslagsværk* vol I, edited by Poul Alstrup, Erik Olsen. Copenhagen: Dansk Historisk Fællesforening, 1991 pp 440-42.
- "Kirkebøgerne" In *Salmonsens Konversations Leksikon vol XIII*, edited by Chr. Blangstrup. Copenhagen:Schultz Forlagsboghandel, 1922.
- Løkke, Anne. *Døden i Barndommen. Spædbørnsdødelighed og moderniseringsprocesser i Danmark 1800-1920*. Copenhagen: Gyldendal, 1998).
- Rosen, Wilhelm von (ed.). *Rigsarkivet og hjælpemidlerne til dets benyttelse I*. 2 vols. Copenhagen: Rigsarkivet, 1983.
- Rosen, Wilhelm von (ed.). *Rigsarkivet og hjælpemidlerne til dets benyttelse II 1848-1990*. 4 vols. København: Rigsarkivet, 1991.
- Ørberg, Paul G. *Hvad præsten skrev - i kirkebogen : kirkebøger og deres brug*. Arkivernes informationsserie. Copenhagen: Rigsarkivet, 1991.

### 3.3 Copenhagen Burial Register

The Copenhagen Burial Register for the period 1861-1942 <sup>7</sup> includes burials that took place in the city of Copenhagen. It was introduced in 1861 by the Copenhagen municipal authorities to centralize the administration of burials in the city. Until then, each graveyard and cemetery registered its own burials locally. From 1880, all graveyards and cemeteries in Copenhagen were managed by a new department, the Copenhagen Burial Authorities (*Kjøbenhavns Begravelsesvæsen*).

The Copenhagen Burial Register consists of bound volumes of pre-printed forms consisting of six (later eight) burials per page (see figure 4). The burials were registered semi-chronologically according to month of burial, and each burial was numbered consecutively, making it evident when pages are missing. There are no known missing volumes.

This source provides complementary coverage for mortality for the city of Copenhagen to that provided by the parish registers (described in the previous section).

#### 3.3.1 Availability of images

The bound volumes are preserved in their entirety and stored at the Copenhagen City Archives (*Københavns Stadsarkiv*). Images of the records are available

---

7. The text of this chapter builds on the documentation created by Copenhagen City Archives and displayed in their webpage (<https://kbharkiv.dk/brug-samlingerne/kilder-paannettet/begravelser-i-koebenhavn/begravelser-1861-og-frem/>) and in other accompanying documentation to the data releases ([Archives and rådstuearkivar](#) 2024). To that comes research articles published by members of the Link-Lives team (Revuelta-Eugercios, Castenbrandt, and Løkke 2022; Ludvigsen, Revuelta-Eugercios, and Løkke 2023)



Figure 4: An example of a pre-printed and completed form from January 1896 in Copenhagen Burial Register, showing six individual burial records.

*Personal information such as names, age, residence, and details on death date and cause can be seen in the handwritten part of the form.*

via the [Københavns Stadsarkiv website](#). They were scanned and transcribed by *Københavns Stadsarkiv* in cooperation with volunteers (see more in next section).

### 3.3.2 Coverage

From 1861–1886, the Copenhagen Burial Register recorded the vast majority of burials in Copenhagen except for the relatively few burials in the military, Roman Catholic and Jewish graveyards. The burial register was meant to record those buried in Copenhagen cemeteries. However, research has found that some people (often Copenhagen residents) appear in the registers despite being buried outside the city. From 1887 to 1994, it was mandatory to include all Copenhagen deaths and burials in the register regardless of which graveyard or cemetery was used.

### 3.3.3 Variables

The pre-printed form for each record is shaped as a short narrative text with space left blank to be filled in by the clerks of the Copenhagen Burial Authorities. No instructions were included but the handwritten contents were filled in systematically and in the same way for all records. The record is written in a kind of shorthand with set abbreviations for parish names and other values.

The burial registers contain the following variables:

- Name of the deceased.

- Information on the parents of children is not required by the form, but it is often written as "son/daughter of" and the name of the father or the mother.
- Age of the deceased given in years, and for infants in months, weeks, days or hours. From 1914 the age is substituted by date of birth.
- Occupation is recorded for the majority of individuals. However, for married women and children the occupation given is often that of the husband or father.
- Place of birth appears for the first time in 1912 and from 1914 the printed forms include a space for it with a relatively small space to fit - just one or two words.
- Death date, as well as burial date and time, given in clear date format.
- Burial place, given as the name of the cemetery or, if outside Copenhagen, the name of the town or parish.
- Residence address at death, given as a street and a street number.
- Place of death or where the body was found, if they are different from the residence address.
- Details about the cost of the burial arrangements.
- Cause of death is believed to be taken from the death certificate (see Revuelta-Eugercios, Castenbrandt, and Løkke ([2022](#)) for more details).



## 4 Transcribed datasets

All the sources included in Link-Lives release 2 have been digitised into machine-readable versions by other projects and institutions. The datasets included as part of the release as “original transcriptions” respect as far as possible the original format in which they were received but they include Link-Lives created ids (the **pa\_id** and the **source\_id**) to ease merging of transcriptions and harmonized versions. There are also some minor additions, which we describe for each dataset.

These datasets are presented as part of the folder “main datasets” and they are characterized by ending in `*_cl.csv` as opposed to `*_std.csv`<sup>8</sup>. Table 5 summarises the main ways in which the origin, structure and aims of these projects have affected the transcription.

---

8. ‘cl’ stands for “clean”, as it is the “clean version” we have received from the data producers after all their own post-processing (not data we have cleaned ourselves). “std” stands for “standard structure” (see full description in section 5)

Table 5: Main characteristics of the projects that created the transcriptions used by Link-Lives in release 2.

|                                       | <b>Censuses</b>                            | <b>Parish register</b>                   | <b>Copenhagen burials</b>                                   |
|---------------------------------------|--|--|---|
| <b>Institution</b>                    | Danish National Archives                   | Ancestry                                 | Copenhagen City Archives                                    |
| <b>Duration of project</b>            | 1992–2025                                  | 2019–2020                                | 2016–2023   |
| <b>Aims</b>                           | Coordination of transcription              | Commercial                               | Make source available                                       |
| <b>Target users</b>                   | Genealogists and researchers               | Genealogists (mostly an American public) | Genealogists and researchers                                |
| <b>Software</b>                       | KIIP, a desktop interface                  | Excel spreadsheets                       | Online interface  |
| <b>Connection image transcription</b> | No   | Yes                                      | Yes   |
| <b>Source timeframe</b>               | 1787–1901*                                 | 1814–1917                                | 1861–1911**   |
| <b>Person records transcribed</b>     | 13,579,776                                 | 50,426,370                               | 307,703   |
| <b>Transcription approach</b>         | Full verbatim with no field interpretation | Verbatim with some interpretation        | Verbatim through drop-down menus, which capture variability |
| <b>All information transcribed</b>    | Yes  | No, selection                            | No, selection   |
| <b>Single/double input</b>            | Single                                     | Single                                   | Single  |
| <b>Proofreading</b>                   | All records on selected sections           | Only as part of quality control          | Systematic search for impossible or invalid values          |

Table 5 continued from previous page

|  | Censuses   | Parish register           | Copenhagen burials  |
|--|--|---------------------------|---|
| <b>Standardisation</b>                               | Limited, as post-processing                                | No                        | Implicit through the drop-down menus, which capture variability |
| <b>Optional fields</b>                               | No   | No                        | No  |
| <b>Transcribers</b>                                  | Danish genealogists (volunteers)                           | East Asian subcontractors | Danish genealogists (volunteers)                                |
| <b>Approximate number of transcribers</b>            | +1000 over the course of the project                       | Unknown                   | +100  |
| <b>Post-processing included in received datasets</b> | Yes, available as part of DDD, but not used for Link-Lives | Unknown                   | No  |

\* Only complete censuses by 2021: 1787, 1801, 1834, 1840, 1845, 1850, 1860, 1880, 1885 Copenhagen and 1901.

\*\* The project has finalized the full transcription of the source 1861–1940, but only the first part up to 1901 is included in Link-Lives.

## 4.1 Censuses

The nineteenth and early twentieth century Danish censuses were transcribed by volunteers at *Rigsarkivet* within what has been known for a long time as the [Danish Demographic Database](#) (*Dansk Demografisk Database*, or DDD). The transcription project started in 1992 as the Source Entry Project (*KildeIndtastnings Projekt*), which was a collaboration between the Danish Data Archive (later included within *Rigsarkivet*), genealogists and researchers that aimed to ensure a coordinated transcription of censuses and parish registers. Over the last few decades it has been commonly known as the Danish Demographic Database, and it has had an independent course from other transcription projects at *Rigsarkivet* via its new Crowdsourcing Portal. During 2025, the old software, databases and procedures were discontinued and all transcription moved to *Rigsarkivets*'s [Crowdsourcing Portal](#).<sup>9</sup> The publications by Clausen and Marker (2000), Clausen (2015), and Marker (2015) provide accounts of the history and the processes carried out by *Dansk Demografisk Database* in the period up to 2015, which we summarize below including an update on the latest developments up to 2025.

In the first few years, copies of the books were loaned to volunteers who transcribed them using a desktop application (the [KIIP program](#)), whose output was comprised of two \*.csv files with data and metadata they then delivered to the archive. Volunteers reserved units with the archive to avoid double transcription, which were given a unit identifier (a **kipnr**). The programme was created and updated by volunteers. This practice continued even after images were made available online in the early 2000s as no online interface was available at the time. The process was coordinated by The Danish Data Archive, an independent archive that eventually became a part of what is today *Rigsarkivet*. A board of relevant stakeholders (called the *Kilde Indtastning Komiteen*, the Source Entry committee) including genealogists, researchers as well as representatives from the archives, was responsible for steering the project. This governing body has continuously overseen the transcription project.

From the beginning, the main two sources of interest have been censuses and parish records. However, there has been a larger interest in the censuses and many have been systematically transcribed. Although volunteers had been able to choose the areas and years they were interested in transcribing first, there has been a general recommendation to focus on one or two particular years to be fully transcribed. The website where the data was made available for search was launched in 1997, and has continued over the years under the name *Dansk Demografisk Database* (see [website](#)) but it is in the process of being phased out.

### 4.1.1 Background to the transcription

Since mid-2010s, there have been additional possibilities to transcribe beyond the KIIP program. *Rigsarkivet* collaborated with two genealogical sites that created online applications to facilitate transcription. For the period December 2015 to May 2018, the now-called *Danske Slægtsforskere* (the association of Danish

---

9. As of writing in November 2025, there are plans for phasing out DDD as an independent webpage.

Genealogists) hosted an online tool in which users could transcribe the 1930 census. This census was given to them in exclusivity so they could manage transcription themselves. From 2013, users were able to use the interface of the Danish Family Search website to transcribe both the 1906 and 1940 censuses, which they also held exclusively. As per the agreement with these operators, the data was originally created in these sites but harvested by *Rigsarkivet* and included in DDD and its website.

After the General Data Protection Regulation, or GDPR, became law in May 2018 along with its Danish implementation, the Data Protection Law (*Databeskyttelseslov*), the scope of the project had to be constrained, given that some of these censuses contained information on individuals still alive in 2018 or still protected by the law (the Danish legislation grants 10 years of posthumous protection). In 2018, the latest census that volunteers could choose to transcribe was 1916 (see “Ændringer i forbindelse med nye EU-regler om datasikkerhed (GDPR)” (2018)).

As of November 2025, all preserved Danish censuses up to and including 1906 are in various stages of transcription, totalling more than 20 million records. Nine nationwide censuses have been transcribed in their entirety: 1787, 1801, 1834, 1845, 1850, 1860, 1880, 1901 along with the Copenhagen census of 1885. Table 10 shows that these censuses contain 13,597,776 records in total, capturing a growth in population from 839,687 in 1787 to 2,468,040 in 1901. At the time of writing, the 1870 census is nearing completion.

The data has been produced and stored in different formats over the years. In its latest form, data from volunteers is imported into a Microsoft SQL database where it is stored according to its geography. This structure was designed to match the structure on the DDD website and to speed up searches by not storing all data together. There is one table for each county (*amt*) which includes variables that can handle all the different columns present in all the censuses; data from all census years is appended here. The county tables correspond to the counties (*amts*) within the artificial stable geography created to organize DDD (see below). This version has some minor additional standardization in the form of a birth year that is used to ease searches on the website. In addition, during this long period, DDD has created and made available standardized versions of some of the fully transcribed censuses, aggregating all counties. These standardizations were made in collaboration with, among others, the [North Atlantic Population Project](#) which involved creating household IDs and standardising main variables, including occupation and position in the household, to a series of hard-coded scripts (Clausen and Marker 2000; Clausen 2015).

Table 6: Number of transcribed records in each census year (from versions delivered to Link-Lives). Note that census of 1885 was carried out only in Copenhagen.

| Census year  | Number of records |
|--------------|-------------------|
| 1787         | 839 687           |
| 1801         | 937 944           |
| 1834         | 1 138 573         |
| 1840         | 1 266 921         |
| 1845         | 1 468 227         |
| 1850         | 1 405 217         |
| 1860         | 1 748 226         |
| 1880         | 1 979 455         |
| 1885         | 327 486           |
| 1901         | 2 468 040         |
| <b>Total</b> | <b>13 579 776</b> |

Simultaneously, the data stored in the county tables has been improved over the years through ad hoc proofreading of parishes but also through continuous micro-improvements as a result of user feedback, so development has been dynamic. The amendments, however, only affect a relatively small number of cases given the scope of the dataset. Until 2024, the process was that an employee at *Rigsarkivet* received the feedback, fixed the issue and uploaded the new information into the system. This process occurred without recording that the data had been updated: the newest version was always deemed more accurate than the previous one. This means that any extraction of the data might have been slightly different from extraction to extraction if new amendments had been put in place, but it was not possible to control the nature of the differences as no systematic releases were ever created. As of writing, the process of phasing out is creating versioned copies of the censuses and it will be possible to track changes over time.

The data used by Link-Lives was extracted by the team by aggregating all the information from the counties for the fully transcribed years. Data from all 10 censuses was extracted in December 2019. However, as 1901 was not fully completed until 3rd September 2021, a new extraction of 1901 was obtained early 2020.<sup>10</sup>

10. The original extraction from 2019 lacked parts of Copenhagen, which were completed in September 2021.

### 4.1.2 Transcription principles

Given the combined interest of researchers and genealogists, the project’s instructions were to transcribe the censuses verbatim, copying exactly what appeared in the source using almost no interpretation. Being faithful to the original would ensure that the data could subsequently be used for multiple purposes. However, not all transcribers adhered to the “verbatim” approach and implemented slight variations in their transcriptions, mostly with the aim of increasing searchability of the information.

An unfortunate consequence of the early origins of the project is that volunteers never established a connection between the transcription and the images on *Rigsarkivet’s Arkivalieronline* image viewer. In the beginning of the project, it was by definition impossible to connect images to transcription as the very first transcriptions preceded the publication of the images online. However, even when the images were available, volunteers did not systematically record the URL of the images they were transcribing, even when the KIIP program had a specific field to capture that information. An ad hoc solution being implemented in the context of the current phasing out has been to identify at least the first image for each parish in *Arkivalieronline*. However, this new information is not part of the datasets that Link-Lives has had access to.

The variables used to capture the information have remained the same over the years. In general, the mapping of the variables from the images of the source to the database is relatively straightforward: most variables in the dataset refer exclusively to a single column in the original source but there are cases of mismatch of content between the standard variables created by DDD and the actual information in the sources. This situation was created by the design of some variables by DDD for coordination purposes and by the process of transcription. A full codebook with all the variables included from the DDD transcription can be found in table 12.1.

The variables that *Dansk Demografisk Database* created to organize the transcription are the following:

- **kipnr**: when allocating units of transcription, the staff broke down the sources in units that would be feasibly transcribed by a person, which tended to be parishes or smaller units within towns. Each distinct transcription unit was given an alphanumeric code and provided geographic attached information.
- These consisted in the variables **Sogn**, **Herred**, **Amt**, **Type** and **Rigsdale** (parish, district, county, type and part of the kingdom). To manage the transcription process, a pragmatic approach to the geography of Denmark was designed to minimise the need to engage with boundary changes over time. This meant creating a DDD artificial geography, which kept the lower administrative unit of the censuses, parishes (*sogne*), as well as the provincial towns (*købstad*), which were largely consistent with the actual census registration geography and the historical geographical units. In some cases, they divided larger census units in pieces (like neighbourhoods). The information on which specific unit the variable **Sogn** hosted was included in

**Type.** Additionally, these “parishes” were nested within districts (*herred*), which were nested within counties (*amt*) as they existed in Denmark in the period before the latest administrative changes in 1972, instead of in their original administrative geographies. This approach created a persistent, necessarily anachronistic but functional structure to deal with the complexity of changing geographical boundaries. This geographic reference prior to 1972 is commonly known as the Trap 5 geography as it follows the 5th version of [Trap](#) series, a compilation of statistical and topographical information about the Kingdom of Denmark, which describes counties, districts and parishes, and was published during the years 1953-1972. Finally, the variable **Rigsdel** included information on whether the data was from the kingdom, one of the duchies or a colonial settlement, as there were also volunteers transcribing those.

- Each individual person in a record was identified by a software-generated **løbenr\_i\_indtastning**. The combination of **kipnr** and **løbenr\_i\_indtastning** created a unique id for DDD.

The variables that were created to capture the information from the censuses and how they have been used, as well as the effect that this design and the practices around it have had on the data is as follows:

- The information from the column *sted* (place) in the original source was transcribed into a number of different columns that varied between and within censuses, and has not been done in a consistent manner. Some transcribers copied all information from the field into one variable, while others separated the information into several different columns. In all censuses, there is a combined variable for address and property title number **matr\_nr\_adresse**, but in some censuses there are also separate variables for both address and title number (**adresse**, **gadenr**, **forhus**, etc). The place information is often only written for the first person of the household, and while some transcribers transcribed this faithfully, others repeated the information for all individuals in the household. Some information is also recorded from the front pages of the census forms, and it is not always clear when information comes from the front page or the census form itself, or if information has been carried from the top of a page down all the households on the page.
- Family number. There is a lot of variation in how the information was recorded in the census sheets in the first place (as we have described under sources in section 3.1) and transcribers did not systematically capture this information in **Husstands\_familienr**. Thus, there is a lot of variation in what appears in this field stemming from these two processes (original registration and transcription). Some transcribers recorded the numbers verbatim but others interpreted them. For instance, some might only have written the family number for the first person in the household just as it was recorded, while another volunteer may have decided to duplicate the number down all the members of that family. When there was not a sequence number in the original source, volunteers often made up one of their own,



either for each census form or a running number for the entire transcription unit. In general, numbers in the hundreds are likely to have been created by volunteers, while lower numbers could be either. Some censuses also had a field for the number of families within the household, which in some cases has been included in this field where transcribers attempted to separate the different families by giving them different numbers or adding “-F1”, “-F2” etc. as suffixes to the family numbers.

- The full name given in a single column in the original census sheet was transcribed into one column (**avn**), without making any interpretations as to which was a given name, surname or maiden name. Volunteers did, however, make slight changes, contrary to the verbatim instructions. We have noted de-reversing the order of the names (where a surname was listed first), giving a family all the same surname, even if the surname only appeared for the first person, or marking with a ditto sign for the subsequent persons. We have also recorded writing abbreviated names in full, e.g. “Christian” instead of “Chr.”, and on occasion abbreviated full names, e.g. “Chr”. instead of “Christian”.
- Place of birth (when it appeared) also respected the one variable approach (**fødested**), without any attempts to distinguish between the different types of geographical units that were reported. This means that expressions such as “parish x, district y, county z”, sometimes separated with commas, coexisted with any other place name in any granularity: a local area, a place by a city, the name of an island, a city, a country, etc. Sometimes, the expression was not consistent with existing administrative units, providing counties or districts that did not exist or not at the time. Volunteers in some cases changed the order of information (e.g. if the county was listed before the parish), or wrote out abbreviations in full, but they did not systematically alter anything.
- The same occurred with occupation and position in the household), that were captured in the same column in the original census forms (until 1901 when they were split into two (see 3.1.6). In some versions of the transcription programme, volunteers were given an additional column to separate the two types of information into, but not in all. The content of those variables was in some years housed under the database variable **ehrv** (occupation) and in other years under **Stiling i husstanden** (household position). For the censuses 1834, 1840, 1845, and 1850, the position in household has been transcribed in **Erhverv**.
- Years and dates were also transcribed verbatim without any attempt to keep them to numeric values. The 3rd March 1890 could therefore be written 3 marts 1890, 03/03/1890 or in any way that the enumerator decided to write it.
- The additional fields in the Copenhagen census of 1885 relating to reproductive information (see section 3.1)(number of children etc) were only tran-

scribed by a few of the volunteers, so this information is not complete for all records in Copenhagen.

- The same is the case for the fields on the quality of housing in Copenhagen in the census of 1880 (recording rent and number of rooms/windows in the home). In the transcription, these variables have also occasionally been mixed up with each other and even with the fields for the address and the number of marriages.
- Until 1870, the census did not provide a separate column for enumerators to record the sex of a person, though the transcription input program included this field. Some of the volunteers left this blank, following the verbatim instructions, but others inputted a value inferred from name and other characteristics from the person represented in the record. Any values in the variable **sex** before 1870 are therefore interpretations rather than true reproductions of what was on the page.
- In the original source, the number of marriages (1st, 2nd, etc.) was only included in the column for marital status on the census forms in 1787 and 1801, but some transcribers have also filled this variable in when transcribing later censuses, possibly with information derived from the position in the household.
- **handicaps** The censuses of 1850, 1860 and 1880 had 5 different columns to record disability, covering different categories of common disabilities: blind, deaf, deaf and mute, deprived of their senses from birth, and deprived of their senses at a later date. In 1901 disabilities were recorded in a single column using the following codes: D= deaf and mute, B=completely blind, A=mentally challenged (*aandsvag*) from birth or early childhood, S= Insane. The chosen column was often filled in with additional details about the circumstances. For these censuses, there was only one variable available for transcription, called "handicaps". Thus transcribers wrote the the exact expression given in any of these columns in the censuses.
- **Boligtælling**: this variable was only available in 1880 and it corresponds to the actual place of residence (*egentlige opholdssted*) information listed on the *Tillægsliste A* template covering those residents of the places whose stay in the parish was only temporary *for de af Stedernes Beboere, hvis Ophold I Sognet kun er midlertidig*. It is sparsely used in the original census and just as sparsely transcribed by the volunteers. The cell assigned for this information in the original template was very small so it was hard to fill with meaningful information.

#### 4.1.3 Known data issues

As the transcription project evolved, new rounds of transcription guidelines were issued, so there are slight inconsistencies in how some variables have been handled. Even though they were not supposed to, volunteers inserted marks and symbols,

such as question marks and exclamation marks, which have no clear meaning for us now.

Additionally, volunteers often added information to improve the usability of the resulting data. For instance, enumerators often used the word "ditto", "do" or marks like single quotations, commas or dashes to indicate the repetition of a value from the previous record. That could be done, for instance, in position in the household to avoid repeating the word *barn* (child) or *tjenestefolk* (servants), but ditto marks can also be found in other variables: sex, civil status, place of birth and even to mark the same surname for different members of the family. The way these were transcribed by the volunteers was not consistent. Some chose to only type the ditto or the specific mark while others chose to write the actual value in brackets alongside it, so that it would be easier for people to use the data without having to look at the previous record.

Individuals on the additional lists (*tillægslister*) (see 3.1.3) were often included in the transcription, although there are some missing. There are, however, no clear indicators that they were additional people as there were no variables to capture them, and in some instances the transcribers have placed them in the family that they think they belong to, and in other cases they will have been listed separately.

These variations in interpretation do not, however, prevent the resulting dataset from being very close to the information in the source.

## 4.2 Parish registers

### 4.2.1 Background to the transcription

The transcription of the Danish parish registers was undertaken by the company Ancestry and mainly covers the period 1814-1917. It includes all events recorded in the Danish parish registers (*kirkebøger*): births, confirmations, marriages, deaths, arrivals and departures. Transcription was carried out using images of the *Kontraministerialbøger* created by Ancestry, as described in 3.2 but also *Rigsarkivet*'s own original black and white images for the period 1892-1917. Some *Kontraministerialbøger* started as early as 1812 but most of them started in 1814, which is generally taken as the date for the official start and the one we have taken for the overall chronology even though there is some data from 1812. The coverage in the datasets provided by Ancestry includes some records before 1814. Our assessment is that their coverage is not systematic enough to warrant investigation before 1814 but a complete analysis is beyond the scope of the project. The coverage ends in 1917 with only a few dozen events recorded afterwards, likely errors in the typing of years.

The full dataset was made available by Ancestry to *Rigsarkivet* through an agreement allowing the latter to use, display and make available the data to researchers under certain conditions (see 2.7) in exchange for supplying Ancestry with a copy of partial data from DDD to use on its website. The Link-Lives use of the data has been covered by both this general agreement but also through an additional agreement with the Link-Lives Consortium as part of the Innovation Fund project.

Ancestry uses the word 'indexing' rather than 'transcription', which is often used to denote that not all the variables have been included. This is the case for its indexing of the parish records in which, for instance, information such as occupation or child illegitimacy was not transcribed, even if present in the original record. This was related to the fact that the transcription was outsourced to Asia, where it was undertaken by non-Danish natives of the language.<sup>11</sup> Ancestry only chose to commit to the transcription of those variables they could obtain a high degree of reliability for, as reported in conversations with *Rigsarkivet*.

*Rigsarkivet* provided some material to Ancestry to help with the transcription, such as name lists, but it was not involved in designing the transcription project in any way and only received an initial example of the transcription followed by the final files.

In the transcription file received from Ancestry in May 2021 and used in this Link-Lives release, each record is matched to an image on Ancestry's website and their own image metadata, which is different again on *Rigsarkivet*'s system. This file is an extraction from the processed version of the original transcriptions created by Ancestry for their own systems. It does not contain all the variables that are displayed on their website and is not organised or structured in a clear chronological, geographic or event sequence.

The file contains all the variables created for all records in a set structure, even though they are not all in use for each event type (e.g. a death date should not appear in a confirmation but the field exists nonetheless). The mapping of the variables from the actual source to the data file is not as straightforward as one might expect. Even though many variables refer exclusively to a single or part column in the original source, there are also many with apparently overlapping content. Here it can be unclear whether they came from the original transcription or some other type of post-processing. We have tried to the extent of our knowledge of the source material and the data to take decisions that take that into consideration.

This Link-Lives release does not contain, by design, this original file received by Ancestry, given the complexity of dealing with its size: it has more than 20 million records and more than 170 variables. Moreover it does not denote which event type each record describes (see 4.2.2). Thus, instead, as a companion to the standardised version of the data we describe in section 5, the "original files" we provide have already been minimally transformed: the data has been split into separate files by event and the information has been transposed from events into person records (see more in 5), with the corresponding Link-Lives unique id numbers (**pa.id** and **source.id**). The original files can, in principle, be supplied upon request, when contacting *Rigsarkivet* for an extraction.

#### 4.2.2 Transcription principles

Ancestry did not make detailed information available on the transcription instructions given to their subcontractors, but by comparing their files with the original sources, and how Ancestry itself displays the data on their website, we were able to ascertain that the transcription approach was mostly verbatim. In this section

---

11. This section is based on the experiences from the Link-Lives team based at *Rigsarkivet*.

we report our findings on our examination of the data and how Ancestry has captured the information in different variables. Our attempt at a codebook for the data is included in table 43.

- Names: there are several variables for capturing the names of the main person in each record (**NamePrefix**, **GivenName**, **Surname**, **NameSuffix**, **GivenNameAlias**, **SurnameAlias**, **MaidenName**). Parsing out the information from the column where it appeared in the censuses, the allocation of names to these variables was a decision made by the transcribers, since there are no clear column or heading divisions in the original records. In many cases, the transcribers ambiguously separated names: people with more than one first name and/or one surname had their names spread over more than two fields (e.g. surnames appearing in **Suffix** or **Alias**, with middle names put into the **Surname** field). All names are treated this way, for each of the additional persons in an event; for example **MotherInLawGivenNameAlias** or **FatherInLawGivenName**. There are variables to contain names for the following roles: fathers (e.g. **FatherGivenName**, **FatherSurname** etc.); mothers (e.g. **MotherNamePrefix**, **MotherGivenName** etc.); spouses (e.g. **SpouseNamePrefix**, **SpouseGivenName**, **SpouseSurname** etc.) and finally mother- and father-in-laws.
- Event types: Ancestry’s transcription includes variables with information about births, baptisms, confirmations, marriages, deaths, burials, arrivals and departures. That means that for both the start and the end of life we have a biological and a religious event. Even though Danish parish registers were created to primarily record vital events (births and deaths) (see section 3.2), it seems the Ancestry data has prioritized the religious events. That means that the baptisms and burials are considered the main event (as in many other European cases). This was evident in an early version of the data provided by Ancestry<sup>12</sup> where a variable event type was present. This approach creates some interpretation issues in relation to the non-Christian registrations, especially the Jewish parish registers, where baptism information contains circumcision date for boys and is not filled in at all for girls. (See section 5.3 for a description on how we have dealt with this in Link-Lives).
- Event type identification: we had access to an early file where this variable was present, but it was unclear to us whether the values had been transcribed by the transcribers or created afterwards by Ancestry based on metadata from images. However, it was not included in the final version of the dataset delivered by Ancestry so we had to create our own (see 5.3).
- Location information associated with the event is available through a set of hierarchical geographical variables for each event type.
  - Location variables are named by using a prefix for the event and a suffix for the type of location, e.g. **BaptismParish**.

---

12. This is also confirmed by the language they use in their webpage, where they have used “dåp” for records, rather than “fødd”.

- \* Event prefixes: **Arrival-**, **Birth-**, **Baptism-**, **Burial-**, **Confirmation-**, **Death-**, **Departure-**, **Marriage-**.
- \* Location suffixes available: **-Place**, **-Parish**, **-City**, **-Municipality**, **-County**, **-Country**. However, not all event types have the same number of variables, with some of them often not including **-Place**, and **-Country** including only Denmark as a default.
- There is also a set for three additional types: **Residence-**, **Vital-** and **Religious-**. The latter two are largely unused.
- The prefixed variables have a substantial number of missing cases (sometimes in the 20-30% range).
- There are two additional variables we believe were created by Ancestry from metadata of their images. They are **BrowseLevel** (county) and **BrowseLevel1** (parish). By comparing the data to what Ancestry has made available on their website, we believe the information comes from metadata that Ancestry attached to each roll of images identified by the **ImageFolder** and/or **SourceReferenceNumber**. This explains the 100% coverage of these variables.
- The date of the event is structured similarly to the location variables.
  - Dates are recorded in variables ending with **-Day**, **-Month**, **-Year**, for all the events listed above.
  - There is a similar variable created by Ancestry, **BrowseLevel2**, which contains the year range of the data, also probably extracted from the images of the original books, and thus describing the year range of the book transcribed.
- Roles of individuals. In Ancestry's transcription, each row in the dataset represents an event (e.g. a confirmation). Information about the main individual in the event (the newborn, the person being confirmed, the groom, the deceased or the departing/arriving person) is stored in columns without any prefixes (e.g. their name would appear in the variable **GivenName**). Additional person information appears in the columns with role-prefixes (**Father-**, **Mother-**, **Spouse-**, **FatherInLaw-**, **MotherInLaw-**). For example, in marriage events, the groom's information will be in the columns without prefixes while the bride's information will be under the columns with the Spouse-prefix.
- In general, information about the additional persons (e.g. a confirmand's or spouse's parents) and secondary information about main persons (e.g. a confirmand's or a spouse's birth date and place) is not placed in separate columns in the original source. This means that such data must have been extracted from the original columns with mixed textual content - through interpretation by the transcriber.
- Chronological coverage: a clear majority of the dataset covers the period 1814-1917 (see table ??). A new parish record law in 1812-1813 forced all

Danish parishes to buy new physical volumes, and this seems to have defined the beginning point of the transcription. Some parishes, especially in the Southern Jutland region where the 1812-1813 law was not implemented, have data transcribed from older books as well. 1917 seems to have defined the end point of transcription, and data from later than this year must be expected to be erroneously dated.

- Image id: each record is connected to an Ancestry image and a publication of the transcription on their website. This can be easily accessed by replacing the image id contained in the variable **ImageFileName** in the URL of Ancestry after "images/": e.g. [https://www.ancestry.se/imageviewer/collections/61607/images/48603\\_324054000599\\_1453-00231](https://www.ancestry.se/imageviewer/collections/61607/images/48603_324054000599_1453-00231). While in principle, *Rigsarkivet* has published the same images, the connection between them has not been re-established.

#### 4.2.3 Known data issues

- Dates of events are not 100% filled: this is very likely due to the transcription process. In many cases, the year of an event was written at the top of the page in the parish register, rather than for each event on the page; or, the year was only recorded once in the register, despite two dates being given, particularly for births and deaths, that have both a date of birth/death and of baptism/burial. This is consistent with the transcriber's choice of the religious date as the marker for the event. For instance, in baptisms, only 72% of the **BirthYear** variable contains data, while 91% of the **BaptismYear** variable contains data. Some cases of non-standard data, such as year-ranges, also appear (e.g. 1892-1895 instead of a single year).
- The dates of birth assigned to persons being confirmed or married are not always correct. Sometimes a vaccination date has been registered in the original source, and this date has been misinterpreted by the transcribers as a birth date (even if it does not fit with the given age).
- In baptisms, the child's last name is sometimes missing. It seems to be the case in a little under 10% of the cases, but the issue is more prevalent in the two first decades of the dataset (up to 35-40%). This is probably mainly due to the original sources and not to the transcription.
- Particularly in marriage and burial registers, the 1891 parish register reform meant that registers after 1892 contain much more information on, for instance, birth dates, birth places, parents/parents-in-law and spouses.
- Locations are inconsistently registered. E.g. place of birth, when registered, seems to have come from the original event entry, but other geographical data might have been interpreted from metadata. As above, for a single event, the information on where the event took place can be found in several sets of variables.
- The roles given to supporting persons are not always correct, e.g. a "father" in the transcription can in some cases be a head of household (*husfader*

= house-father) or the name of an unwed mother’s father rather than a biological father. There is more certainty about familial connections after the 1891 reform.

- Variables can contain characters (random spaces or text strings) that are not accurate reflections of the original image. Often this happens in columns that are not relevant for the particular event, e.g. -InLaws columns for burials.
- There are a number of duplicate registrations since many events were originally registered twice. E.g. a person who died in one parish and was buried in another can be recorded in the registers of both parishes. The full extent of this issue is unknown; it is not an issue arising from the transcription itself but from the nature of the source.
- Information about stillbirths was recorded in a number of different variables in both the births and deaths variables (**BaptismAge**, **DeathAge**, **BurialAge**, **Notes**).
- Ages in marriages can be problematic. We have tested small random samples of the transcription against the original images and found missing ages in the transcriptions as well as transcription errors where ages were wrong by several years. There is no consistent pattern to these mistakes, although both errors and omissions seem to be considerably higher in some parishes than in others. This suggests that it is a result of varying precision among the transcribers. The full extent of this issue is unknown.
- Mothers’ ages in birth events are inconsistently transcribed. They have been transcribed in the dataset from the 1850s onwards but cease to be transcribed around 1892, even though the original sources continued to record the information.

### 4.3 Copenhagen Burial Register

The Copenhagen Burial Register 1861-1940 was transcribed by volunteers at the Copenhagen City Archives (*Københavns Stadsarkiv*) in a crowd-sourcing project that started in 2016.<sup>13</sup> Link-Lives used the records from 1861-1911, which were all those fully transcribed at the time of delivery to Link-Lives. By August 2023, the transcription of all records in the Copenhagen Burial Register for the period 1912-1940 was complete. Volunteers transcribed the records via *Københavns Stadsarkiv*’s website, using an online web application managed by the archive. *Københavns Stadsarkiv* provided detailed instructions for the transcription (see below).

---

13. As in section 3.3, the text of this chapter builds on the documentation created by Copenhagen City Archives and displayed in their webpage (<https://kbharkiv.dk/brug-samlingerne/kilder-paa-nettet/begravelser-i-koebenhavn/begravelser-1861-og-frem/>) and previous work by members of the Link-Lives team (Revuelta-Eugercios, Castenbrandt, and Løkke 2022; Ludvigsen, Revuelta-Eugercios, and Løkke 2023)



As described in section 2.2, the copy of the data that is incorporated in this Link-Lives release dates from March 2021 and was delivered to us in this format. At the time, there were some years after 1911 that had already been transcribed but we chose to focus on the period to 1911. An improved version of the transcription has been made available by Copenhagen City Archives but unfortunately we have not been able to include it. As volunteers have improved the quality of the existing transcriptions and extended the coverage up to 1940, there are inconsistencies between the version that can be found on the website and the dataset used in our Link-Lives v.2 data release. But as we have kept ids and URL connections, users can check those for themselves.

The data actually starts with 49 deaths in December 1860 and contains a death from 1842 recorded in the year 1862. Additionally, there are two deaths from 1854 and 1856, which are the result of transcription mistakes (the transcriber wrote the birthdate instead of the deathdate) and should be discarded.

Table 7: Number of valid records included in Link-Lives version 2 from the Copenhagen Burial Register divided by decade.

| Decade  | Number of records |
|---------|-------------------|
| 1860-69 | 32227             |
| 1870-79 | 46215             |
| 1880-89 | 65474             |
| 1890-99 | 71681             |
| 1900-09 | 75736             |
| 1910-11 | 15503             |

#### 4.3.1 Transcription principles

The overall principle in this transcription project was that the transcription should be true to the original source.<sup>14</sup> While transcribers were allowed to cross-reference other sources (e.g. by locating the deceased in the Copenhagen parish registers) to help disambiguate unclear or difficult handwriting, they were instructed to only transcribe what was in the burial protocol itself and to not “correct” what they might perceive as errors.

Many of the fields in the transcription interface had available drop-down lists of values (in a slightly standardised format) that the volunteer had to find in the corresponding printed field in the original burial register. Transcriptions using these lists are still true to the original source in the sense that the precise wording

14. This section is mostly an English translation of the instructions in Danish to volunteers, which can be accessed at <https://kbharkiv.dk/brug-samlingerne/kilder-paa-nettet/begravelser-i-koebenhavn/indtastningen-af-de-kobenhavnske-begravelsesprotokoller-1861-1940>/<https://kbharkiv.dk/brug-samlingerne/kilder-paa-nettet/begravelser-i-koebenhavn/indtastningen-af-de-kobenhavnske-begravelsesprotokoller-1861-1940/>.

is generally preserved, except that unambiguous abbreviations are written out, typos are corrected, and spelling variations have been standardised according to current Danish orthography. The transcribers were instructed to only choose options from the lists that exactly matched what was written in the source. If the correct option did not yet exist in the list, they were to choose “*\*skal oprettes*” (“\*must be created”). To ensure a close match between the transcriptions and the original source, each list was managed and continuously updated with new entries by either an employee at *Københavns Stadsarkiv* or by one of the “superusers”. The superusers are a select group of very experienced volunteer transcribers, with an in-depth practical knowledge of both the original source material and the needs and wishes of the volunteer community. They work closely with the archive to improve the quality of the transcriptions. As part of that they have each been trusted with a number of additional responsibilities, including quality checking the transcriptions, mentoring new transcribers, and managing one or more of the standardised lists according to their own area of expertise. A complete codebook for the dataset, largely derived from the original created by **Københavns Stadsarkivet** is provided in table 44). The variables have been affected by this approach to transcription in the following ways:

- The name of the deceased is divided into first names (**firstnames**) , last name (**lastname** and maiden name (**birthname**. For young children, the last name is often inferred from the recorded name of their parent(s). Names that did not fit into these categories (e.g. nicknames or the first names of parents to dead children) were transcribed in the **comment** field. Some special cases where no names were recorded for the deceased were marked as follows:
  - **Firstnames**: “*Udøbt*” = unnamed infants, “*Dødfødt*” = infants recorded as stillborn, “*Abort*” = burial of a miscarried foetus, “*uoplyst*” = other cases where no first name was recorded.
  - Last name: “*uoplyst*” = no last name recorded. The names of women and children from one of the hospitals who supplied anonymous details were often just recorded as a patient number, and are entered with the first name “*uoplyst*” and the last name “*Anonym nr. XXX*”. If the surname is recorded, it is written as “[surname], nr. XXX”.
- The age of the deceased was transcribed in different columns, depending on whether it was recorded in years, months, weeks, days, or hours in the original record (**ageYears**, **ageMonth**, **ageWeeks**, **ageDays**, **ageHours**). This ensured the age was transcribed exactly as it was recorded, without converting between time units. This means that the variables were only filled when the information appeared in these units. There are no values in all the other variables if the age was given in years. The only exception was the ages of infants of under 1 year: when the information was provided in months, days, weeks, or hours for children under 1, the variable **ageYears** was always to be filled with 0 as a supplement, to make explicit that the record was that of an infant.

- Dates of death (**dateOfDeath**) and dates of birth (**dateOfBirth**) (recorded from 1913) were transcribed in the format dd-mm-yyyy, where the year of death was usually found at the top of the page. If either the day or the month were missing from the recorded birth date, these were replaced with a 1 in the transcription. E.g. if the birth date is recorded as March 1912, it would be transcribed as “1-3-1912”. If no birth date was recorded, the variable was to be left empty.
- Place of death/discovery was transcribed using a partially standardised drop-down list of common places of death or of discovery of bodies. This included “at home” as well as a large number of lakes, canals, parks, public places, hospitals, and institutions. When the recorded place was too specific or detailed to be matched to an individual item in the list, the closest match from a number of broader categories was to be selected. This is often the case when the record recorded the circumstances of how and where a body was discovered, or when the place of death was a specific address, on board a named ship, or in a location that could only be specified with a lengthy description. If a broad category was chosen, the precise string from the original source was transcribed in the **comment** column as either “*Dødssted*” (precise place of death) or “*Findested*” (precise place of discovery).
- Civil status (**civilstatus**) was chosen from a fixed list containing the standardised options for marital status as prescribed in the pre-printed instructions in the original source. If none of the printed options were marked in the original documents, the civil status could sometimes be transcribed from a status/title added to the name (e.g. “Unmarried Betty Petersen”). The standardised list therefore also includes options for children who were explicitly recorded as legitimate, illegitimate, orphans or foundlings. See an example of the values available in table 8.
- The sex of the deceased (**sex**) was interpreted by the transcribers based on other recorded information such as name and occupation, with “unknown” as an option when the deceased could not be identified as either male or female.
- General information about the place of burial (cemetery, chapel, parish) was transcribed by choosing the exact match from the drop-down lists.
- The residence/address of the deceased person was transcribed in several different columns, depending on how it was recorded in the original source (**street**, **parish**, **hood**, **street\_unique**, **street\_number**, **letter**, **floor**, **institution**, **institution\_street**, **institution\_hood**, **institution\_street\_unique**, **institution\_street\_number**):
  - If the residence was an address in Copenhagen, each component was entered in separate fields: street name (drop-down list, with neighbourhood included to distinguish between multiple streets with the same name), house number and letter (separate columns, both free text transcriptions), and the floor/apartment (drop-down list).

Table 8: The standardised value of **civilstatus** in Danish and their respective translations in English.

| Civilstatus            | English translation                                       |
|------------------------|---|
| Ugift                  | Single  |
| Gift                   | Married   |
| Enkestand              | Widowed   |
| Uægte barn             | Illegitimate child  |
| Vides ikke (ukendt, ?) | Original source notes civilstatus as uncertain or unknown |
| Skilt                  | Divorced  |
| Separeret              | Separated   |
| Ægtefødt barn          | Legitimate child  |
| Forladt                | Abandoned (usually by spouse)                             |
| Hittebarn              | Foundling   |
| Forældreløs            | Orphan  |

- For addresses outside of Copenhagen, Frederiksberg or Gentofte were transcribed verbatim, as free text, in a separate variable called **addressOutsideCph**.
- If the residence of the deceased was recorded as an institution or a location other than an ordinary address, its name was chosen from the drop-down list in the field **Institution**. This includes hospitals, charitable housing associations, prisons, barracks, hotels, famous or public buildings in Copenhagen etc. plus some broader categories to be used in the same way as described above for the place of death/discovery.
- The transcription interface allowed users to select multiple causes of death from a drop-down list. The transcribers were instructed to enter each cause of death separately, in the order they were recorded in the original source. The individual causes are separated with commas in the **deathcauses** field.
- The transcription of occupation/position (either recorded in the **position** field in the record or as a title in front of a name) also allowed for multiple entries per burial. The occupation/position can belong both to the deceased and close relatives, and a single burial can contain the records of occupations for several different persons. The relationship between the deceased and each of the transcribed positions (own current position, own former position, position of the spouse or parent etc.) were entered in a separate column in the transcription. In the dataset, the order of the occupations in the

**position** column corresponds to the order of the comma-separated entries in the **relationtypes** column.

- The **comment** field was included to compensate for the limitations in using the fixed drop-down menu lists. If the transcriber chose an entry from the standardised list in one of the fields that was not a precise match for the original string (e.g. “American Ambassador in transit” would be occupation=“Ambassador”), the precise transcription was to be entered as a comment, in the format [field name]:[exact transcription of the original string in that field]. Some transcribers also looked up the burial in the parish registers out of their own interest and might then have included a reference to relevant records in the comment field.
- The transcription includes the variable **id** that connects to the exact image from where it was transcribed and the most updated transcribed data at [Københavns Stadsarkiv](https://kbharkiv.dk/brug-samlingerne/soeg-i-indtastede-kilder/post/1-56160). In the viewer, replacing the last numbers after the “1-” with the value of **id** from the transcribed version of the dataset in the URL <https://kbharkiv.dk/brug-samlingerne/soeg-i-indtastede-kilder/post/1-56160> yields the desired page for a given **id**.

#### 4.3.2 Known data issues

- Not all entries in the burial registers were records of a new burial. Apart from recording who was buried where and when, the records also documented any payments received in relation to each funeral, although this part of the form was not transcribed as part of the project. If the relatives of the deceased later wished to pay additional services, f.x. maintenance of the grave-site or to keep the plot in the cemetery beyond the usual period of time allotted to each grave, the transactions were sometimes recorded on a new form in among the actual burials in the record, with only the name of the deceased and the details of the burial filled in. Although the volunteers were instructed to skip these accounting entries, some have occasionally been transcribed by mistake.
- Some burials contain multiple deceased persons. This sometimes occurs if siblings, especially twins, died at the same time, or if mothers and infants both died during childbirth. In such cases the names were both to be transcribed in full, separated with “&”. For the other variables, especially date of birth, place of birth, and cause of death, it is not specified in the original source which of the deceased individuals the information belongs to, though it can sometimes be deduced from the context.

## 5 Link-Lives data harmonisation

We use the word harmonisation to denote the process by which we create new versions of the original datasets in the format developed by Link-Lives, consisting of new variables derived from the original values. These are called the harmonized datasets and they are structured and named similarly across all datasets in the project (see codebook in section 12.4). They are characterized by ending in `*_std.csv` (standard structure) as opposed to `*_cl.csv` (clean version from the data providers) (see also section 4). These datasets are used for linking and can make analysis easier. We highlight variable names in **bold** to distinguish from concepts, for instance we refer to the variable **age** but talk about issues about age in general.

The overall principle for harmonisation has been to maintain a single data structure that can absorb different types of transcriptions based on what we call a "person-appearance", where one record in the dataset corresponds to the information of one individual appearing in a source in a place at a given time. In the standardised datasets, the variables are named with generic terms in English that capture different potential dimensions of an individual's source appearance; for instance, age, sex, place of birth. These variables do not contain the original transcribed values but standardised versions. For instance, the variable **age** in the standardised dataset for the 1880 census does not have the actual content of the variable from the transcribed dataset, which is a text field containing the value "39", for example. The variable **age** in the standardised dataset is a numeric interpretation of that variable: an integer.

In order to fit this structure, the original variables have been treated in a number of ways according to the specificity of the original transcriptions and the needs of both linking and analysis. In the next sections, we discuss the general approach towards harmonising the datasets and specify special procedures when relevant to the different types of original transcriptions.

### 5.1 Establishing source IDs

In order to easily and unequivocally identify the different source datasets in Link-Lives, we have assigned them codes, as in the following table 9.

### 5.2 Identifying person appearances

As described in the introduction to this section, one record in the dataset corresponds to the information of one individual appearing in a source in a place at a given time. Link-Lives has created a unique **pa\_id** (person-appearance ID) for each individual record in the standardised dataset, in addition to any unique IDs the original dataset may have contained. That means that a **pa\_id** is only a unique identifier within a given source and that the combination of **pa\_id** and **source\_id** is necessary to uniquely identify a person appearance across all Link-Lives datasets.

For transcribed datasets like the parish records that include information on more than one individual in the same record, we split each record into as many

Table 9: Source ids values used in Link-Lives.

| Source.id | Source name                  |
|-----------|------------------------------|
| 0         | Census-1787                  |
| 1         | Census-1801                  |
| 2         | Census-1834                  |
| 3         | Census-1840                  |
| 4         | Census-1845                  |
| 5         | Census-1850                  |
| 6         | Census-1860                  |
| 7         | Census-1880                  |
| 8         | Census-1885                  |
| 9         | Census-1901                  |
| 10        | Copenhagen Burial Register   |
| 11        | Parish register-burial       |
| 12        | Parish register-baptism      |
| 13        | Parish register-marriage     |
| 14        | Parish register-confirmation |
| 15        | Parish register-departure    |
| 16        | Parish register-arrival      |

records as there were named individuals to create the standardised dataset. This procedure significantly increased the number of actual records in the dataset, that now correspond to person-appearances instead of the original event appearances. The information for each of these individuals (generally kept in separate variables, such as **person1\_name**, **person1\_surname**, etc.) was mapped onto the standardised variables for names, sex, etc. in the corresponding record. In order to keep track of this splitting process, we performed a series of procedures:

- A new **event\_id** variable was created 1 to 1 to the original id of the record.
- **pa\_ids** are created for each of the individuals mentioned in the original record.
- A variable **role** records the assigned role of the individual within the record, according to the particularity of the source.
- The information from the variables **event\_type**, **event\_time** and **event\_place** is also replicated in each person appearance (see below).

### 5.2.1 Censuses and Copenhagen Burial Registers

The structure of the census and Copenhagen burials datasets already fits the person-appearance format, so each person in these datasets was given a Link-Lives **pa\_id** and the information from the original values is easily mapped to the standardised variables. Technically, the Copenhagen burials can also be considered events where more than one person could be mentioned, in the form of a limited amount of double burials and through recording information about next of kin (usually last name + occupation/title of spouse or parent). However, due to the reduced number of cases with that information and the way it was recorded, it has not been feasible to split the information into multiple records as we did for the parish records. However, not all records present in the data were converted to person appearances. A limited amount of records in the censuses and Copenhagen burials were left out because they did not have values for the year of death. The number of records in the final file is shown in table 10.

Table 10: Number of harmonized records in censuses and Copenhagen burials

| Source             | Original records | Harmonized records) |
|--------------------|------------------|---------------------|
| 1787               | 839687           | 839687              |
| 1801               | 937944           | 937944              |
| 1834               | 1138573          | 1138573             |
| 1840               | 1266921          | 1266921             |
| 1845               | 1468227          | 1468227             |
| 1850               | 1405217          | 1405217             |
| 1860               | 1748226          | 1748226             |
| 1880               | 1979455          | 1979455             |
| 1885               | 327486           | 327486              |
| 1901               | 2468040          | 2468040             |
| Copenhagen Burials | 307703           | 306839              |

### 5.2.2 Parish registers

The transcribed parish registers were supplied to us in a format in which each record contained an event (birth, marriage, death, etc.) and there were multiple variables for name, age, place of birth for each type of individual (FatherGiven-Name, for example). The standardised Link-Lives dataset for parish registers follows the general description of the procedure in section 5.2. The information about each of the individuals present in a given event record (e.g. parents and child in a baptism record) was mapped into the standardised variables for names, sex, etc. for each person record in the standardised structure. The split



is achieved by first always capturing the information of a main person (though sometimes there is no name for the main person, e.g. in case of a still born child) and then identifying names registered in the specific name variables for other roles (“Father”, “Mother”, “FatherInLaw”, etc). If a name was found in one of the role-specific name variables, an additional **pa\_id** was added with the corresponding role and the same **event\_id** as the main person with the relevant person or role specific variable values. See section 4.2 for a list of all potential variables with names for different roles and table 11 for the results. Any additional variables associated to these named roles were also transferred to the corresponding person appearance.

Table 11: Total number of individual person appearances in the parish registers (harmonized dataset), divided by decade and event type.

| Decade                | Arrival | Baptism  | Burial  | Confirmation | Departure | Marriage | All      |
|-----------------------|---------|----------|---------|--------------|-----------|----------|----------|
| 1800-09               | 940     | 50207    | 46352   | 4594         | 248       | 23525    | 125866   |
| 1810-19               | 53815   | 489972   | 258331  | 292097       | 46351     | 151164   | 1291730  |
| 1820-29               | 176138  | 920996   | 481709  | 516513       | 164530    | 242481   | 2502367  |
| 1830-39               | 391366  | 1056939  | 581623  | 679998       | 386823    | 259301   | 3356050  |
| 1840-49               | 573374  | 1199208  | 583583  | 706693       | 581674    | 287213   | 3931745  |
| 1850-59               | 605408  | 1414316  | 649735  | 809333       | 615770    | 377859   | 4472421  |
| 1860-69               | 594615  | 1581498  | 709688  | 904457       | 612840    | 369776   | 4772874  |
| 1870-79               | 298527  | 1797899  | 741148  | 997942       | 312123    | 408268   | 4555907  |
| 1880-89               | 1152    | 2001163  | 785832  | 1055514      | 2774      | 404839   | 4251274  |
| 1890-99               | 786     | 2262937  | 1293750 | 1401132      | 196       | 930968   | 5889769  |
| 1900-1909             | 453     | 2485405  | 1401449 | 1586294      | 578       | 1209829  | 6684008  |
| 1910-19               | 2       | 1790143  | 1046305 | 1333020      | 0         | 956062   | 5125532  |
| 1920-29               | 0       | 8        | 25      | 61           | 0         | 0        | 94       |
| Uncertain/<br>unknown | 247855  | 2065284  | 266246  | 495148       | 249352    | 142848   | 3466733  |
| All                   | 2944431 | 19115975 | 8845776 | 10782796     | 2973259   | 5764133  | 50426370 |

### 5.3 Event\_type

This variable describes the type of event a dataset contains and can consist of the following values: census, baptisms, marriages, burials, arrivals, departures, confirmations, burial protocols.

While the values for censuses and Copenhagen burials are evident from the nature of the sources, we reconstructed event types for the parish registries given that there was not a variable explicitly included in the Ancestry extraction (see 4.2.2).

We have used the way Ancestry named its variables to create events. The name of the variables including information on dates and locations started with an “event” (Birth, Marriage, etc...) as well as information about the individuals they refer to (Mother, Father, etc). We created an algorithm that investigated what type of information was recorded to assign a type of event. In order to decide on the main event, we followed Ancestry’s sub-optimal approach to centre parish register types around the religious instead of the biological events (baptisms instead of births, burials instead of deaths). This meant to ensure that we

did not tamper with the internal consistency of the transcription and to maximise the information present at the event level.<sup>15</sup> However, this solution posed other challenges. In both Ancestry’s and our standardization, Jewish births have a baptism date instead of a birth date which obviously distorts the actual information of the source. We hope to improve event identification in the future with solutions that can both allow us to maximize the information available and better account for the diverse nature of the sources. As of December of 2025, we have adjusted the way the events are named in linklives.dk and we hope to change it in the final release.

As a result of this procedure, there were 3,704 records to which we could not assign an event. This was mainly due to transcription errors in records with special layouts or characteristics.<sup>16</sup> See table 11 for the breakdown of person-appearances and events as a result of the conversion to Link-Lives standardised format.

## 5.4 Role

In order to account for the different roles held by a named person in an event, we created a variable called **role**, which defines how a person-appearance is related to an event. Table 12 describes the main roles used for the different types of events or datasets. There are some cases where some events also have person-appearances with roles not typical of that type of event, e.g. baptisms where we find mothers-in-law or fathers-in-law. This mostly reflects either errors from the Ancestry transcription or the allocation of event type by the automated process within Link-Lives.

## 5.5 Event date variables

There are several variables that provide information about the date of an event. An **event\_date**, was standardised into either an aggregated format in one variable, or into three variables separated into day, month and year. This was to capture the variability in how event dates could be recorded and resulted in three variables: **event\_year**, **event\_month**, **event\_day**.

### 5.5.1 Census

We pre-filled all four variables with the date on which the censuses took place. The date has been derived from the original source.<sup>17</sup>

---

15. For instance the proportion of records which have a birth year was lower than those with a baptism date (72 vs 91%). So choosing baptism as an event allowed us to be able to use more information

16. These ambiguous records are allocated to a separate .csv file, that is not a part of this release.

17. For 1787 it was 1 July, for 1801, 1 February, for 1834 18 February, for censuses in the period 1840-1921 it was 1 February

Table 12: Roles used in the Link-Lives standardised data according to event and dataset.

| Events and dataset | Roles used  |
|--------------------|---|
| Censuses           | (set to missing)  |
| CBP Burial         | deceased  |
| Baptism            | baptised, mother, father  |
| Burial             | deceased, father, mother, spouse  |
| Confirmation       | confirmand, mother, father  |
| Marriage           | groom, bride, bride-mother, bride-father,<br>groom-mother, groom-father |
| Arrival            | arrived, spouse, mother, father   |
| Departure          | departured, spouse, mother, father<br>(fatherinlaw, motherinlaw)        |

### 5.5.2 Parish registers

We converted all event-specific date variables (for year, month and day) with a prefix for an event (for baptisms, marriages, burials, arrivals, departures and confirmations) into event variables for each type of event and created the relevant variables. As the transformation involves making them into date or integers, we have only worked with variables recorded as integers/dates in the Ancestry transcription. There is a large potential for improvement in completing this task. **event\_date** has values only if **event\_year**, **event\_month**, and **event\_day** have all content. Incomplete date information means that there is a substantial number of missing cases.

We have treated similarly the other two event variables that complement the religious baptism and burials events: **birth dates** and **death dates**.

Birth dates and death dates appear mostly in baptisms and burials, respectively. There are three original variables for each of them (**BirthDay**, **BirthMonth**, **BirthYear** and **DeathDay**, **DeathMonth**, **DeathYear**). They are mostly directly coerced into integer values, the exception being that a synonym catalogue is used on the **BirthMonth** variable to convert the various ways months can be noted using non integer values. If the **BirthDay** variable contains a value larger than 31 it is replaced by an empty value. Everything is then converted into a date-time-format. Anything that cannot be coerced into an integer value is replaced by an empty value.

There is an additional set of date variables in the original data called Vital (**VitalDay**, **VitalMonth**, **VitalYear**) which in some cases seems to have been used instead of the prefixed event variables during the transcription process. However, we have not yet determined whether and/or how best to include that

information into event dates.

### 5.5.3 Copenhagen burial register

The date chosen is that of the person's death, which is recorded in the variable **dateOfDeath** (yyyy-mm-dd). Its values are transferred to **event\_date** and also divided into day, month and year for the respective variables.

## 5.6 Age/birthdate

### 5.6.1 Census

In **age**, we have coerced all values from the transcription into numeric values for the censuses where ages were given (before 1901). We have deleted ages over 115 and applied a minimal set of procedures to extract numeric values out of expressions. Anything that cannot be recognised is replaced by an empty value. For age in 1901 we have calculated the age as a difference between the census date and the given date of birth.

There is no accounting for the differences in age recording between the pre- and post-1870 censuses, where age is given differently (see section 3.1 on age recording).

F

As in 1901 the birthdate was given (YYMMDD), we transformed the dates, stored in string format, into separate variables and then concatenated them into a proper date unless the resulting age did not fall within a realistic range, i.e. negative numbers and numbers above 115. Only cases where the day, month and year were available were calculated.

### 5.6.2 Parish registers

We have constructed an **age** out of the age variables for baptisms, marriages, burials, confirmations, departures and arrivals. We have performed a ordinary cleaning of the **EventAge** variable, which includes removal of non-numeric characters, conversion of fractions, etc. See section above on event dates for birth year information. The variable **SpouseMarriageAge** is a combination of event and role and is mapped to **Age**.

### 5.6.3 Copenhagen burial registers

As age was already stored in 5 different variables (years, months, weeks, days and hours), we have taken the information expressed in years to fill in **age**. Except when **ageYears** was 0, then the standardised age is calculated as a decimal number using age in months, weeks, days, or hours.

**Birth\_year**: the standardised birth year was taken from the original CBP variable **dateOfBirth** when it was filled out or **yearOfBirth**. If neither were filled out, the standardised **birth\_year** is calculated from age and **event\_year**.

## 5.7 Event\_place

We have 6 variables that account for different levels of location. The lowest unit is contained in either **event\_location**, if available, where a specific place within an administrative unit was given but it is not always present. However, each record was assigned to an **event\_parish** (*sogn*) and, if part of a town (*købstad*) in **event\_town** with the overall name of the town, that often aggregated several parishes. In addition, the parishes and/or towns were placed in their corresponding higher units, districts (*herred*) in **event\_district** and counties (*amt*) in **event\_county**. Additionally, we have **event\_county**, which is filled by default with the value "Denmark".

### 5.7.1 Census

These variables contain the information about where the person resided in the census year and are closely related to the original variables from the transcription *sogn* (parish), *herred* (district), and *amt* (county), which have its own special geography (see 4.1). The main change has been to make them lowercase and replace the old spelling "aa" for "å".

- **Event\_county**: the direct value of the *amt*.
- **Event\_district**: the value of *herred* with minor special changes, as replacing *København (Staden)* for *København*.
- **Event\_town**:
  - Left empty when the unit is a parish.
  - Filled with the value *København* when the content of *Sogn* is a street in Copenhagen.
  - Filled with the name of the town present in *Sogn* when the unit is actually a town (Type = *Købstad*) but removing the word *købstad*.
- **Event\_parish**: the value for rural parishes (i.e. Type=*Sogn* or Type=*Delsogn* in the original transcription– the last is usually rural districts connected to towns, such as *Ringsted Landsogn*) is standardised to the value of the persistent geography created by Link-Lives for a parish (see section 9.3.1 for more details). For Copenhagen and provincial towns, **Sogn** includes street names or neighbourhoods but we have not transfer this information to any other variable.

There is additional information in the census variable **Stednavn** that could indicate lower hierarchy units but the content is very heterogeneous so there has been made no effort to integrate it in **event\_location**.

No attempt has been made to standardise other address-level location information (in **Sogn** or any of the other variables that may contain addresses, **address**, **gadenr**, **matrikel**, etc. as it is not consistent enough for our purposes. This is of particular relevance in the towns and cities.

### 5.7.2 Parish registers

The relevant event-specific place variables (parish, county) with a prefix of an event (baptisms, marriages, burial) have been harmonised and standardised to create **event\_parity** and **event\_county**. However, There is additional information on places in the variables **BrowseLevel** and **BrowseLevel1** as well as in the Vital set and we are working on improvements for next releases.

Our current procedure does the following:

- Cleans the event place variables for each event, especially parish and county.
- Checks that these combinations exists in the reference list of parishes in our permanent geography and that the parish and county match each other.
- It stores the results in the variables **event\_parity** and **event\_county**.

This process is far from complete at the moment of the release of version 2: there are many missing cases, standardisation is sub-optimal and when **event\_county** is not matched to a real county, we can not provide a parish. We are working on improving this procedure on several fronts.

### 5.7.3 Copenhagen Burial Register

All records have been assigned “København” as **event\_town** based on the collection, purely as “metadata”.

## 5.8 Birth places

In this section we describe only the “reported” birth places from e.g. censuses or death registrations. That is, birth places which have in common that they were reported later in life than the birth itself. The birth place of the birth registrations is handled in section 5.7. The main approach to birth place was developed for censuses and applied to the rest of the datasets very early in the project. It handles some of the main variation but it still requires additional developments. Especially the advent of Large Language Model holds a lot of promise to improve the quality and coverage of this variable.

### 5.8.1 Census

As places of birth were recorded in the original sources in a text field (see 3.1.6), we first extracted all unique, original place name strings (over 600,000) from the transcribed census datasets 1845-1901 and split out into individual words (e.g. “*Ølstrup Sogn, Ringkøbing Amt*” became “*Ølstrup*”, “*Sogn*”, “*Ringkøbing*”, and *Amt*). This resulted in a list of more than 130,000 unique words, the majority of which appeared only once or very few times. For efficiency reasons, only the most frequently appearing strings were then manually standardised: 5,000 name strings, providing information on c.97% of all the unique expressions in the censuses (but less than 1% of words). These words were coded by a historian with expert domain knowledge of the original sources and linking. Key words such as *Sogn* or “S” (for *sogn*) and *Amt* or “A” (for *Amt*) appearing in the original

strings were used for classifying the unique place words as names of parishes, counties, etc.

The standardisation was made to a reference set of words, with accompanying information about the type of places that were associated to that word, created from a combination of the Danish Demographic Database geography and DigDag. The domain experts used a special software tool created for the purpose, where the words that had an exact match with a reference word/place were pre-standardised. The domain-experts could also select some words as “keywords” if they were not expressions denoting place but used around it.

The classifier/parser looked at specific combinations of words and keywords with different types of patterns where all words were recognised.

- When the string had one word (30%), it tried to match it to our standard list. And if the word could refer to different types of units (i.e., Copenhagen could be a town and a county), we assigned the type of place following this order: town, County, Parish, Place, no class.
- When the string contained two words and it matched a given pattern (7%): “nameparish namecounty”, “namecounty *amt*”, “namecity *køgstad*”, “nameparish *sogn*”.
- When the string contained three words and it matches a given pattern (6%): “nameparish *sogn* county”, “nameparish namecounty *amt*”, “*koebstad* nametown namecounty”, “*koebstad* namecounty *amt*”.
- When the string contained four words and it matches a given pattern (20%): “nameparish *sogn* namecounty *amt*”, “*koebstad* nametown namecounty *amt*”.
- *Her i sogn* (15%). If string matched “*her i sogn*”, “*i sogn*”, or “*født i sogn*” the values for **birth\_parity**, **birth\_county** etc. were filled with the values of the corresponding **event\_place** variables

### 5.8.2 Parish registers

We used the original **BirthPlace** variable as input to version 2 of the above mentioned birth place algorithm. There is ongoing work to improve this variable. The **BirthPlace** variable is not treated for the Arrival and Departure event types.

### 5.8.3 Copenhagen burials

This variable is not available for the current period.

## 5.9 Names

### 5.9.1 Census

Our harmonization approach was created for the 1845-1901 censuses, which contain the most complex data for this variable: the full name is reported as a single variable with no classification into first names, family names, etc. Before classification we perform a cleaning and standardisation:

**5.9.1.1 Cleaning** We remove all non-alphabetic characters except the dot “.” including anything written inside square brackets “[]” which is where transcribers usually added additional information, for instance with the values of dittos.

**5.9.1.2 Standardisation** We standardise the names using a synonym catalogue containing the 5432 most frequent names and surnames. This synonym catalogue was created by taking the full name strings of all censuses available at the time of creating the catalogue (1787-1901) and splitting them into single names based on a **space** separation, i.e. “Hans Kristian Jensen” is split into “Hans”, “Kristian”, and “Jensen”. Then a frequency count of each name is performed. We then chose to standardise the most common names which account for 95% of all name occurrences in that set of censuses. (A full account of the construction of the synonym catalogue is provide in section 9.2, the simplified file with the synonym catalogue is provided as part of the release (`SC_names_v1.csv`), whose structure is described in 12.10).

**5.9.1.3 Classification** The classification of the names largely follows the method by Wisselgren et al. (2014) and classifies them into first names, family names, patronymics, true patronymics, maiden family names, and maiden patronymics. Below we describe some of the issues we encounter.

1. We refer to the groups of names not including the first names as last names.
2. Maiden name: In some records the maiden name is explicitly marked by the word “*født*” (born). Names after “*født*” are only categorised as a maiden name if there is a first name and the sex is female.
3. After removing the information about maiden names, we continue by classifying the remaining names into first names, family names and patronymics.
4. During the period some name traditions changed gradually. Specifically we see a transition from giving girls a true patronymic, e.g. the daughter of *Niels* would be named *Nielsdatter* or *Niels Datter*, to using the patronymic as a family name, i.e *Nielsen* (Niels’ son) instead. There are many examples of switching between the two systems in the course of a life, so that a *Nielsdatter* might later go by the name *Nielsen*. In an attempt not to miss any links and to avoid erroneous links because no other good candidate links could be found, we standardise all patronymic suffixes to “*sen*”, e.g. *Nielsdatter* is standardised to *Nielsen*.
5. Some records for children do not explicitly note the last names. In those cases we try to construct a set of possible last names if we identified a father in the household. The family names and patronymics of the father are then transferred to the child accordingly. We also consider the possibility that the child was given a true patronymic which we construct from the first name of the father.
6. Likewise, for some married women the last name was not noted. In some cases, however, the married women were recorded with their maiden names



(without explicitly marking them as such) but had already or would later adopt or use the last name of their husband. For any married woman not sharing any last name with her husband we also consider that information.

**5.9.1.4 Resulting variables** Along with the variable **name\_cl** (which includes the full name in lower case) and the variable **name** (where all the elements have been standardised), the names are split into 5 variables used for linking: **first\_names**, **patronyms**, **family\_names**, **maiden\_names** and **uncat\_names** (uncategorised names). This last variable contained elements that are not/cannot be standardised. The unstandardised name components are, thus, missing from all steps that come after cleaning the raw name strings into **name\_cl**. This results in cases of totally missing or partial name strings in the variable **name**. The unstandardised name components are missing from all steps that comes after cleaning the raw name strings into **name\_cl**.

We also have two additional variables which include both actual names and patronymics as well as other derived from husband/father to create potential additional names to match for those who do not have one in any other of the first 3 variables. These are **all\_family\_names** and **all\_patronyms**.

## 5.9.2 Parish registers

The splitting up of the event-based records in the transcribed dataset into a number of person appearance records in the standardised dataset includes relocating the rolespecific name variables (mothers, spouses, etc) to the relevant person appearance and **pa\_id**. The list of name variables available from the original data for each named role is as follows:

- **GivenName**
- **Surname**
- **MaidenName** (Does not exist for Father and FatherInlaw roles)
- **GivenNameAlias**
- **SurnameAlias**
- **NamePrefix** (Does not exist for Father and FatherInlaw roles)
- **NameSuffix**

**GivenName**, **Surname**, and **MaidenName** are cleaned and stored in **name\_cl**. Each of the three variables are standardised according to the regular synonym catalogues and concatenated into a single variable, **name**. Following this approach, **name** is used as input to the usual name categorisation algorithm (see above). The standardised **MaidenName** is separately categorised into patronyms and family names to produce maiden patronyms and maiden family names, which are used in **all\_family\_names** and **all\_patronyms**. All “*datter*” suffixes on patronyms are replaced by “*sen*”, and all possible last names are gathered into two variables, dividing them in all possible patronyms and all possible family names. Note that

**Aliases**, **NamePrefix**, and **NameSuffix** are not included in our current name standardization.

### 5.9.3 Copenhagen burial registers

We concatenate and clean all variable names (**firstnames**, **lastname**, and **birth-name** into **name\_cl**, which was then the basis of the standardisation and categorisation in the same way as for census and parish registers.

## 5.10 Sex

Link-Lives use “male” and “female” as main values for sex. There are currently no values to distinguish between empty values in the original source and typos or other markings. See table 13 for actual values used in the data.

Table 13: Codebook for sex in synonym catalogue (standard values).

| standard | description |
|----------|-------------|
| m        | male        |
| f        | female      |

### 5.10.1 Census

For censuses where the variable **sex** was not recorded on the original form, transcribers inferred a value in 60% of cases. To achieve a higher coverage, a neural network was employed for predicting an individual’s sex based on their name and household position. The neural network was trained using samples where transcribers had previously identified the **sex**. Transcribers had access to various details, including name, household position, civil status, contextual household information, and occasionally occupation and other pertinent details. We trained the neural network using a harmonized version of sex derived from a basic synonym catalogue we created. The process involved creating a list of original values by first taking all the values from sex from the censuses, cleaning alphanumeric symbols, trimming spaces and converting them to lower case. Afterwards, our domain experts went through the highest frequency occurrences and standardized them. The file used is provided as part of the release 2 **SC\_sex\_v1.csv** and its structure is described in section 12.10. Values not present in the synonym catalogue have been set to missing.

The assignment of the most probable value for **sex** was generally straightforward due to this rich set of information. The training dataset encompassed sex information for over 4.6 million individuals, with no fewer than 480,000 samples for any given census. Using this data, the frequency of classification as male or female was calculated for each name and household position. These statistics were then transformed into features for every name and household position, enabling the neural network to predict sex. The obtained results demonstrate high

reliability, with a precision score ranging between 99.2% and 99.7%. In order to ensure comparability, the value in **sex** reflects, for all years, the value of the neural network and not a standardisation.

### 5.10.2 Parish registers

The variable **sex** is cleaned, removing any non-relevant characters (hyphens, spaces, etc) and then standardised according to the simple synonym catalogue described in section 5.10.1. Any other value is set to missing value (NaN). **Sex** of main person in event in births and deaths is derived from division into men and women in the parish records. When events are split into person appearances, parents are assigned a gender according to their role, i.e. mothers = female, and fathers = male. There is no explicit record of the sex for the roles “Father”, “Mother”, “FatherInLaw”, “MotherInLaw”, so one is inferred from the prefix when the events are transformed into person appearances and given a **pa.id**. For marriages, the role of “Spouse” for **sex** can be recorded but is often not. In those cases **sex** is inferred as the opposite sex of the main person. Arrivals and Departures only very rarely have a transcribed **sex**, so neither they nor their spouses have a standardised sex.

### 5.10.3 Copenhagen burial registers

The original Danish values “*mand*” and “*kvinde*” have been translated to the standard in English. The value “*ukendt*” has been standardised as a missing value (NaN)

## 5.11 Civil status

Civil status is standardised to a limited set of values, that can be seen in table 14. Other potential values which do not specify a marital status (separated, etc) have been set to missing (NaN) for simplicity purposes.

Table 14: Values of marital status used in Danish and their translation to English.

| Values | Translation |
|--------|-------------|
| ugift  | single      |
| gift   | married     |
| skilt  | divorced    |
| enke   | widowed     |

### 5.11.1 Census

We performed a simple harmonization to Link-Lives standard values using a synonym catalogue. The synonym catalogue was created by first taking all the

values from marital status from the censuses, cleaning alphanumeric symbols, trimming spaces and converting them to lower case. Our domain experts went through the highest frequency occurrences and standardized them. The file used is provided as part of the release 2 `SC_marital_status_v1.csv` and its structure is described in section [12.10](#). Values not present in the synonym catalogue have been set to missing.

### **5.11.2 Parish registers**

The variable was often not filled as there was no direct information provided on civil status in the source.

### **5.11.3 Copenhagen burial registers**

We performed a simple harmonization to Link-Lives values using the synonym catalogue created for censuses (see section [5.11.1](#)). Values not present in the synonym catalogue have been set to missing.

## 6 Linking methods

Link-Lives v.2 contains two sets of nationwide links between our sources created through two automatic linking methods: 1) rule-based (see section 6.2) and 2) machine learning (see section 6.3). We also provide a benchmark dataset created through computer-assisted domain expertise record linkage (see section 6.1) on a sample of records. The creation of this dataset has fulfilled a dual aim: acquire sufficient knowledge about the process of record linkage from a historian’s perspective and acquire sufficient training and testing data to support automatic method development. In this section, we first describe the domain-expert method for record linkage followed by the automatic methods that use it for training and/or testing. For these three methods, we provide link rates as summary measures of performance, accompanied by a summary quality assessment.

### 6.1 Computer-assisted domain-expert record linkage

We take as point of departure the generalized consensus on two issues related to data linked by humans: it represents the gold standard in record linkage<sup>18</sup> but it does not constitute a “ground truth” (Abramitzky et al. 2019; Bailey et al. 2017; Akgün et al. 2020). Any record linkage performed between two records that were not originally connected is indeed a construction, no matter how likely it seems to be “true”. In addition, we also take into consideration the scholarship on human behaviour, especially from economics, on the human subjectivity and inconsistency in systematic decision-making, as described in the work of Kahneman (2013), that we find very relevant to the task of human linking.

The literature generally describes the training data used for developing automatic methods as “manually linked data” or “hand-linked data”. However, we have coined the term “computer-assisted domain-expert record linkage” to name our own approach for creating “manually linked data”, to highlight the fact that it is a method that requires explanation. It is aimed at ensuring transparent, consistent and reliable results. Its key features are: reliance on trained domain experts to find the most likely pairs of links within a set of guidelines unconstrained by computer-generated pre-selection, measures to reduce unwanted variability in human judgment, the use of a computer interface to support and facilitate linkage; and restriction of data used for linkage to the same information that will be accessible to algorithms. The latter is particularly important to highlight as it means that we do not strive to achieve the best possible linkage results for a given sample of data, which will require a genealogical/historical approach, cross-checking of multiple sources. Instead, we create the best estimates that can be produced with the data that algorithms will be able to use.

The result of the implementation of the method in the project has produced what we call “the benchmark dataset on linked censuses, parish registers and Copenhagen Burials, 1845-1901”, used to test models described in section 6.2 and train and tests models described in section 6.3. However, we have also applied and

---

18. It is important to note that there are as many approaches to record linkage as there are projects, with examples of approaches that do not involve any form of human validation or linked data, as in Mandemakers et al. (2023)

expanded for testing the earlier period (“Benchmark dataset on linked censuses, 1787-1845”).<sup>19</sup> It is available in the file `benchmark_v1.xlsx` within the `links` and `lifecourses` folder.

In this section, we describe briefly the main elements of the method, that is, all the different research decisions taken to define the linking process following our framework based around three questions: “what we link” (section 6.1.2.1), “how we link” (section 6.1.3) and “who links” (section 6.1.4). At the end, we provide a summary of the overall link rates that it has produced for the main period 1845-1901 and the extension 1787-1901 (section 6.1.5). Additionally, section 10 provides additional details on the specific implementation of the method to create the benchmark dataset for the period 1845-1901, used for both training and testing. Section 11 describes the implementation of the method for the 1787-1845 benchmark dataset. Furthermore, the additional documentation included as `paradata`<sup>20</sup> includes a variety of working documents used during the project, to further provide a transparent account of the conditions under which data was created (see description of what is `paradata` in section 2.4).

Our approach has been grounded in the specific set of sources that we had to link, as transcribed datasets. The decisions were initially tied to the censuses, our main sources, but we adjusted the approach as the project included new sources, developing specific variants of the overall approach while keeping the spirit and comparability with the decisions taken for censuses. Table 15 outlines the main features of the slightly different variants of the method for each of the source pairings.

Developing the method has required taking decisions on three main areas: what we wanted to link, how we were going to do it and who was going to do it. Answering these questions has meant making decisions about a variety of technical and human aspects. As part of the process, we have developed an interface, a series of data processing protocols and pipelines and accompanying documentation. In the following sub-sections, we describe them in detail.

---

19. This data has not been used for the purposes of testing in this release

20. `Link-Lives Paradata.Computer Assisted Domain Expert Linkage.pdf`

Table 15: Main features and characteristics of methods in domain-expert linking of different sources.

| DIMENSIONS   | Submethod 1  | Submethod 2                                      | Submethod 3   | Submethod 4  | Submethod 5  | Submethod 6  |
|--|--|--|---|--|--|--|
| WHAT TO LINK: OVERALL DECISIONS IN THE LINKING PROJECT |  |  |   |  |  |  |
| Aim  | Training and test  |  |   | Test data  |  |  |
| Sources  | Census to census   | PR baptisms to census                            | PR marriages to census                                      | PR burials to census                                 | CPH burials to census                                | Oldest census to oldest census                               |
| Period   | 1845, 1850, 1860, 1880, 1901                                 | All available                                    |   |  |  | 1787, 1801, 1834, 1840                                       |
| Sequence of linking                                    | One-to-one, sequentially                                     | One-to-one, PR to the closest census             |   |  |  | One-to-one, sequentially                                     |
| Target units for linking                               | Parishes for rural areas, samples of streets for urban areas | Parish   |   |  | Sample of consecutive individuals for selected years | Parishes for rural areas, samples of streets for urban areas |
| Sampling of units                                      | To maximise chronological and geographical representation    |  |   |  | To maximise chronological coverage                   | To maximise chronological and geographical representation    |
| Sampling of individuals within units                   | Full coverage  | Father, mother and child compulsory. Rest of kin | Bride and groom compulsory. Rest of kin only if co-resident | Deceased compulsory. Rest of kin only if co-resident | Full coverage  |  |

Table 15 continued from previous page

| DIMENSIONS   | Submethod 1  | Submethod 2  | Submethod 3 | Submethod 4 | Submethod 5 | Submethod 6 |
|--|--|--|-------------|-------------|-------------|-------------|
| HOW TO LINK: SPECIFIC RULES & RESTRICTIONS FOLLOWED IN LINKING |  |  |             |             |             |             |
| Versions of source allowed to be consulted                     | Transcribed and standardised   |  |             |             |             |             |
| Access to source-external information                          | Geography dictionaries   |  |             |             |             |             |
| Variables used   | Invariable + household context   | Invariable + household context + occupation and residence as secondary                             |             |             |             |             |
| Software assistance  | Potential candidates created by algorithm and general search                 |  |             |             |             |             |
| Possible linking outcomes                                      | Link, maybe, multiple, not found, unborn                                     | Link, maybe, multiple, not found, unborn, maybe+secondary, multiple+secondary, not found+secondary |             |             |             |             |
| Direction of linking   | Backwards  | Forward  | Forward     | Backwards   | Backwards   | Backwards   |
| Iterations   | One  | Some   | One         | One         | One         | One         |
| WHO LINKS? Dealing with human variations in judgement          |  |  |             |             |             |             |
| Number of linkers  | 2  |  |             |             |             |             |
| Adjudication method  | Linker feedback on initial disagreements + arbiter on enduring disagreements |  |             |             |             |             |
| Linker profile   | Domain experts in history  |  |             |             |             |             |



Table 15 continued from previous page

| DIMENSIONS                    | Submethod 1  | Submethod 2 | Submethod 3 | Submethod 4 | Submethod 5 | Submethod 6 |
|-------------------------------|--|-------------|-------------|-------------|-------------|-------------|
| Linker<br>training/guidelines | Initial training + written guidelines + ongoing refreshments |             |             |             |             |             |

### 6.1.1 Purpose

The main purpose of developing the method was to create training and test data. However, given the limited amount of existing systematically linked datasets in Denmark<sup>21</sup>, we also wanted to create sufficiently representative data for minor local studies that could be used to additionally test the performance of automatic methods on limited research questions. However, in one occasion we applied virtually the same method to create a test dataset (e.g. submethod 6 described in table 15). The only difference between their implementation has been in the size and sampling of the units to ensure representative test sets.

### 6.1.2 What we link? Linking scope

#### 6.1.2.1 Sequence

We have taken a sequential pair-wise approach, linking only two sources at a time. We have established censuses as the backbone of the project, so first we have linked all the censuses sequentially, one to one, and then the rest of the sources to the relevant censuses. We have not had resources to implement the internal linking of parish records (for instance, births and deaths) or between parish records and Copenhagen Burials.

#### 6.1.2.2 Target units and sampling strategy

Rather than drawing random samples of individuals or household, we focused on larger areas so we could get a sense of how a particular area was linked in its totality. We defined the geographical units that we aimed to link as “linking units”, which in most cases consisted of full parishes. In urban areas, we sometimes created samples of streets or neighbourhoods (see section 10 for full details).

We have aimed to select geographic units that were socially diverse and geographically spread across Denmark. Chronologically, we have selected units representing all time periods within our sources. We have linked them to their nearest available census counterpart (e.g. a burial registered in 1875 would be linked backwards to the 1860 census).

To create a dataset exclusively aimed at testing for the period 1787-1945 (see section 11), we have sampled linking units of c. 250 person appearances aiming for a location split of 50% rural, 30% town and 20% Copenhagen records, ensuring that all persons sharing the same household ID (i.e. all householders) were included.

#### 6.1.2.3 Sampling of individuals within linking units

---

21. There are several works by Johansen, which used the family reconstitution method in the early days of historical demography (Johansen 1975) but very few examples afterwards. Among them, we have to highlight Thomsen (2010), whose linking approach, using all types of sources in three parishes in Jutland, and Thomsen’s expertise as member of the project has been an inspiration to the development of the nuances of method for Danish sources.

We sampled individuals within our selected sampled areas according to slightly different criteria depending on the specifics of the source:

- Census-to-census: we attempted to link every person in the household using their own registered information.
- Parish records: as multiple individuals were typically registered within the same event, we chose to attempt to link only what we have termed “key-individuals”: the child and parents in a baptism, the married couple in a marriage and the deceased in the burials. If other relatives were mentioned in the record, the link was only made if they could immediately be spotted while performing the link of the key person, that is, in the same household. We called this approach “linking of convenience”, as it was biased towards stable family co-residence. The reason for creating these additional links was to gather additional qualitative work that could be used in automatic approaches<sup>22</sup>

### 6.1.3 How we link? Specific choices in Linking method

#### 6.1.3.1 Data allowed for linking

We have only allowed the use of information that our algorithm could also use (or we consider that we could incorporate it in the future), to minimize introducing noise in the models for training or testing, setting unrealistic expectations of what the model would be able to link. Thus, we have used exclusively the information provided by the two sources that were to be linked in its transcribed format received by the project or with minimal standardization (see 9.1). We also have not allowed for cross-checking with other project sources, the actual images or any other external data, as would be typical in a family reconstitution project or genealogical work. However, we have acknowledged one exception to external material: geographic resources on-line. Given that additional geographic information could be leveraged systematically in the future, we have allowed searching and identifying historical place names using two designated websites. We are aware that there is a gap between the life-courses that a full domain-expert method (a genealogical-historian approach) can reconstruct for a given individual and what our domain-expert approach can achieve with limited tools and data. The best approximation to full reconstruction of individuals life-courses still requires domain-expertise and use of a variety of sources, some of them not available in machine-readable format, triangulation between them and several iterations. Thus, we on purpose produce sub-optimal data from the perspective of full life-course reconstruction but optimal for model training and testing purposes.

#### 6.1.3.2 Choice of variables

---

22. This approach maximizes the number of links overall but obviously biases link rates downward and skews representative analyses if we take all individuals present in records. So these links have not been used to calculate link rates or any other calculation unless specified.

Since Ruggles seminal work, there has been long an ongoing discussion on which variables can and should be used for linking (Ruggles 2002). Following that discussion, we divided variables in two groups (primary or secondary) according to how they could be used to establish whether two records may belong to the same individual.

- Primary variables: the classic invariable information proposed by Ruggles (Ruggles 2002, Ruggles and Magnuson 2020) plus the household context, which is also now accepted as containing relatively acceptable biases (Antonie et al. 2014, Thorvaldsen, Andersen, and Sommerseth 2015 and Helgertz et al. 2022): that means birth year/birth date, birth place, gender, consistent civil status, plus household context.
- Secondary variables, we have included occupation and place of residence.

For most of our census linking, we have only used primary information<sup>23</sup>. But we have used secondary variables for linking Copenhagen burials and parish registers to censuses in a design fully comparable with the census approach through a two-step approach.

- Domain-experts first took a decision on whether a record from source A could be linked to another from source B based on primary information, without looking at secondary variables.
- Afterwards, they used secondary variables to add a second decision if the secondary information is used.
- The final button selected by the domain expert registered the two steps, which allows us to account for the known biases that secondary variables may introduce into the linked data.

This approach helps to disambiguate cases with collision of multiple candidates with similar invariable characteristics but different residence or occupation. As these two sequential decisions (on primary and secondary variables) have been recorded explicitly as metadata on every single link, we can afterwards create two versions of the benchmark dataset, one conservative, fully compliant with the exclusive use of primary variables (if desired by the researcher) and an extended dataset, with decisions derived from considering also the secondary information.

### 6.1.3.3 Software assistance

We developed a program called ALA (Assisted Linkage Application), tailored to our needs given that off-the shelf regular tools, as MS Access or Excel, could not accommodate the size of our datasets or offer sufficient ease of use to implement our decisions. It is a standalone locally-stored application, programmed in Kivy and Python, which we could zip and circulate among our staff.<sup>24</sup> Figure 5 shows

23. We only used occupation for linking censuses 1901-1880 and 1880-1860 but we specifically marked those decisions (see button option under the description of our software)

24. We expect to release the software that can be used with the data included in this release during 2026. Contact us at data@rigsarkivet.dk if you would like to have access before that.

a screen-capture of the main screen of ALA. The interface allows users to link a linking unit (a parish from a census, parish records, Copenhagen burials or a user-generated dataset) in the left side of the screen) to the corresponding full national census (on the right side). The program includes several features that aid linking:

- A subset of potential candidates generated by a basic rule-based algorithm, using some relatively simple rules on matching names and year of birth (see specifics on ALA user manual) that appear when the user clicks on any of the records to be linked (middle right).
- A series of search boxes to manually refine searches outside of the potential cases proposed, where it is possible to use a variety of wildcard search (top right).
- The full household for individuals selected for examination (as a result of either choosing a pre-defined potential candidate or the result of the use of the search boxes (down right)).
- A full comparison of individual information for the two records being compared, as not all the information is displayed on the main panels (bottom-left).

Figure 5: The linking interface (ALA) with options to browse data, search the sources and make link-decisions.

The screenshot displays the ALA (Automated Linkage Assistant) interface. It features a main table of potential links with columns for R.no., id, Link, Name, Birth place, Res. parish, Res. county, Birth year, and Score. Below this table are several panels for detailed comparison and decision-making.

**Top Panel: Search and Filter**

| Name             | Birth year       | Birth place   | Sex                   | Marital st. |
|------------------|------------------|---------------|-----------------------|-------------|
| Occ. information | Res. information | Res. parish   | Res. county           | id          |
| Name 2           | Birth year 2     | Birth place 2 | Household proximity 2 | Clear       |

**Middle Panel: Potential Links**

| R.no. | id      | Link | Name           | Birth place                  | Res. parish  | Res. county | Birth year | Score |
|-------|---------|------|----------------|------------------------------|--------------|-------------|------------|-------|
| 77    | 1359762 | -    | Kjeld Sørensen | Bælum S. (Sogn Aalborg Amt)  | Søbyeg       | Aalborg     | 1785       | 0.0   |
| 112   | 1349017 | -    | Kjeld Sørensen | Her i Sognet                 | Klarup       | Aalborg     | 1833       | 0.0   |
| 25    | 1351556 | -    | Kjeld Sørensen | Dtø (Her i Sognet)           | Romdrup      | Aalborg     | 1835       | 0.0   |
| 234   | 1245698 | -    | Kjeld Sørensen | I Sognet                     | Gullev       | Viborg      | 1799       | 0.0   |
| 305   | 171413  | -    | Kjeld Sørensen | Sørum Sogn                   | Sørum        | Hjørring    | 1806       | 0.0   |
| 642   | 123011  | -    | Kjeld Sørensen | I Sognet                     | Vensted      | Hjørring    | 1826       | 0.0   |
| 230   | 278675  | -    | Kjeld Sørensen | Rindø                        | Bøtø Kræfter | København   | 1822       | 0.0   |
| 449   | 1198249 | -    | Kjeld Sørensen | Her i Sognet                 | Brande       | Vejle       | 1848       | 0.0   |
| 176   | 688488  | -    | Kjeld Sørensen | Mølbacke Sogn (Bjerskov Amt) | Trøde        | Brøndby     | 1814       | 0.0   |

**Bottom Left Panel: Full person info 1878-1890**

| Link                   | ImageFolder                  | ImageFileName | id   | SequenceNumber | PageSeq |
|------------------------|------------------------------|---------------|------|----------------|---------|
| 48438.22021000034.2082 | 48438.22021000034.2082.00007 | 47832         | 1000 | 211            |         |

**Bottom Right Panel: Full person info 1850**

| Link                     | Place  | Res. parish | Res. county    | Res. information | Occ. information |
|--------------------------|--------|-------------|----------------|------------------|------------------|
| Fureby, Belgum, Hjørring | Fureby | Hjørring    | Huus, Mølgaard | Almåslem         |                  |

**Bottom Center Panel: Additional personal info 1850**

| R.no. | id     | Res. information | Place                    |
|-------|--------|------------------|--------------------------|
| 452   | 157830 | Huus, Mølgaard   | Fureby, Belgum, Hjørring |

There are a series of buttons for recording the outcome for a given linking attempt in a more nuanced way than just a yes/no decision based on a given set of variables. The buttons allow for combining (1) a graduation of choices between a link and a not link and (2) the information that have been used for taking that decision.

(1) The graduation of choices between a link and not found consist on five options:

- We have allowed two negative options, “not found” or “multiple”, to differentiate cases where no candidates were really feasible (“not found”) from cases of collision, with a few equally similar good candidates from those where there is just not any good candidate (“multiple”).
- We have defined “link” and “maybe” as two positive linking outcomes to allow domain experts to document their own level of certainty. This reflects the early experiences in the project that linking records is rarely a clear, binary yes/no decision and to allow the psychological safety of recording some level of uncertainty that could later potentially be accessed. (See section 10 for a summary of our best practices on how we have operationalized the distinctions).
- We have identified as “unborn” those individuals who could not have been born at the time of the census according to their age and thus cannot be “not found”.

(2) The variables used to take a decision are registered through the use of an additional “+ secondary button”. For any decisions but for “unborn”, the button ribbon allows to additionally register a second decision using the button “+ secondary”, that indicates the use of information from the fields of either occupation or address to confirm the likelihood of a match.<sup>25</sup> The use of the buttons and their relationship is clearly described in a set of guidelines (see section 10).

#### 6.1.3.4 Direction of linking

As most sources can be linked backward or forward and we had limited resources, we chose to focus on the direction that maximized the likelihood of finding persons in the target source. This decision led to different decisions for different source pairs.

- Census to census: backward to the earlier counterpart (e.g. 1860 to 1850). This ensured that people would be alive and we did not need to consider the effect of mortality.
- Burials (from parish registers or Copenhagen): only meaningful backwards.
- Baptisms and marriages: forward in time to the next available census. Given the paucity of information available for each individual, it was easier to find the groups of individuals (married couples or couples with a child) together co-residing in the census (e.g. a marriage in 1879 was linked to 1880, not 1860).

---

25. For a period during our census to census linking, as a test while we considered the need to include support information, we created a button “link with occupation” that was used to mark that particular case. See its introduction in table 39.

### 6.1.3.5 Iterations

We have limited the linking approach to just one iteration over the full linking unit. Multiple linking rounds could capture additional links once multiple candidates were disambiguated by the first round and increase link rates. However, we decided that it was not cost-effective for us, as the number of new links acquired would not warrant its cost.

### 6.1.4 Who links? Choice of linkers

We refer to “linkers” as the domain-experts who perform linking under our method. The decisions taken under this section have been aimed to cost-effectively reduce variability and ensure consistency in our linked dataset by harnessing the strengths and weaknesses of human judgment. Our specific approach has been inspired by a combination of approaches found in the historical record linkage literature (Bailey et al. 2017; Feigenbaum 2016) and state of current psychological research on human variation in judgment (Kahneman 2013).

#### 6.1.4.1 Number of linkers and adjudication methods

We have defined a process that involves three domain-expert linkers in a process inspired by the approach used by Bailey et al. (2017).<sup>26</sup> It has three steps:

- Independent linking: two linkers link independently the same unit.
- Comparison: we compare the results of both linkers and flag substantial disagreements (i.e., “maybe vs. link” or “multiple vs. not found” is not considered a substantial disagreement). See full list in table ???. The decisions agreed by both are termed “uncontested”.
- Linker review: each linker is asked to revise only the cases where there is a disagreement in the decision and decide whether they maintain their original decision or accept the other decision. They also have to provide a minimal written explanation of their choice. At that time they are allowed to use the interface and attempt the link again. The decisions are labeled according to this step as:
  - Moderate disagreements: when linkers agree in the second phase
  - Enduring disagreements: when they still disagree
- Arbitration: enduring disagreements are sent to a third linker that takes all the material and explanations produced by the first two domain-experts and can also look at the case from scratch in order to make a final decision.

---

26. We have further expanded and tested results by our own work on Danish sources in a series of experiments and found out that this approach produces roughly the same results as having three full independent linkers and taking a majority vote.

The benchmark dataset in Link-Lives release 2 includes information about the process for each single linking attempt in each linking unit (who attempted to link it -identified with a linker id-, what decision was taken and the results of the review and adjudication, as recorded as metadata, including version of interface, timestamps and other details). Further details about the implementation of the process are described in section 10.8.

#### 6.1.4.2 Choice and training of linkers

We decided to select only domain-experts for the task of computer-assisted domain expert linking as we called them “linkers”. We allowed both professionals or domain-experts in training, that is, individuals with prior and active interest/training in historical demography and 19th century Danish social history, largely historians or students in History to carry out the linking. We designed a training course (Linking School) inspired by the procedure described by Bailey et al. (2017) to ensure aligned aims, procedures and results, which everyone needed to take before starting to produce for the team. Additionally, we created a series of “Check-in” events over the course of the project to ensure that consistency maintained over time and that adjustments to new types of sources were also consistently applied. For instance, we carried out linker workshops with some frequency, later replaced by ad hoc sessions for new sources. See more details of these practices in section 10 and specific examples of the documentation used in Link-Lives Paradata.

#### 6.1.5 Results

The results of the implementation of our method is provided in the file “benchmark\_v1.xlsx”, where both data for the period 1787-1845 and 1845-1901 are combined. We have attempted to link 60,680 person appearances in total (58,142 for the main period 1845-1901 and 2,538 for the earlier period). We provide a summary below of the key link rates for both periods.

Link rates are calculated as

$$Linkrate = \frac{positivelinks}{mainlinkable}, \quad (1)$$

where *positivelinks* is the sum of the positive decisions, through primary information (“link” or “maybe”) or also using gsecondary information (“maybe secondary”, “multiple secondary”, “not found secondary”) and *mainlinkable* is the number of individuals that were alive to be linked (removing “unborns”) from the individuals defined as “key persons”. “Key persons” are the selected individuals for a given event that we actively tried to link. In censuses, all persons are “key persons” but in parish registers, we have made selections (see Selection of individuals within sampling units in section 6.1.2.3).

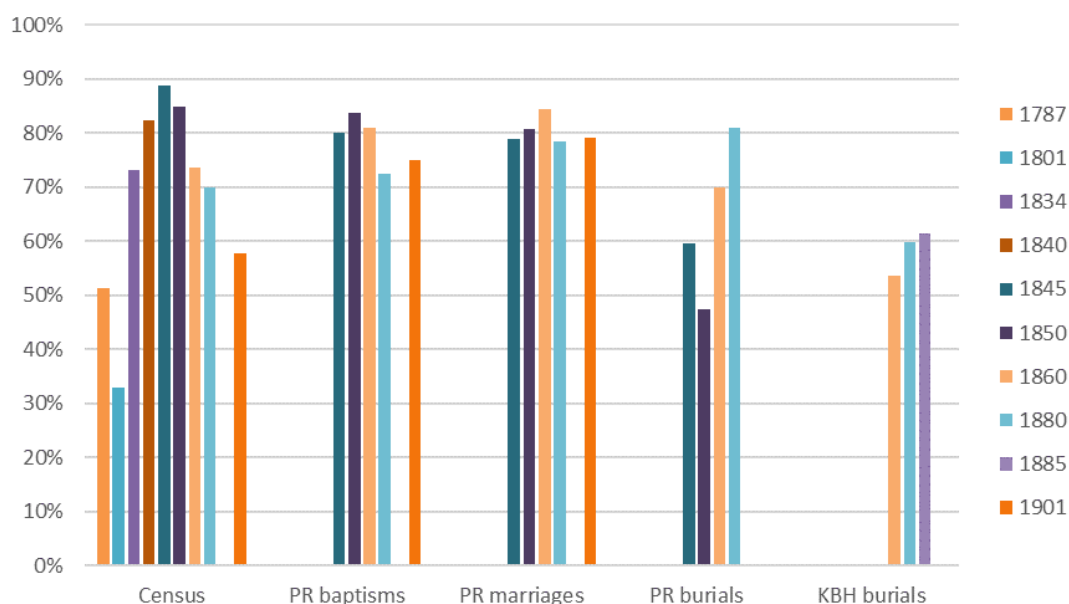
##### 6.1.5.1 Benchmark dataset 1845-1901

We applied the method for the period 1845-1901 for the purposes of training and testing automatic record, with at total of 47,100 records with a recorded decision. We describe additional specific details regarding its construction in section 10).



Figure 6 shows that link rates are higher for census to census pairs and census to marriages and census to births pairs, with rates at around 80%. For burials, either from parish records or from Copenhagen Burial Register, they are at a lower range. There are clear differences in linking rates over the entire chronology caused by a variety of factors specific to the context, linking and the sources, but also related to the number of years elapsed between the two sources. We discuss these findings in a forthcoming article (Revuelta-Eugercios et al., n.d.). One of them is the amount of information available in both types of burials, where there is a tendency to only show information about the deceased, compared to the other sources where more individuals are present. Both the qualitative experience of our linkers and these figures highlight the importance of context and family relationships in providing additional points of information to make linking possible.

Figure 6: Link-rates of the Benchmark dataset 1787-1845 and 1845-1901.



#### 6.1.5.2 Benchmark dataset 1787-1845

In addition, we created a limited benchmark dataset for the period 1787-1845 only on census to census pairings, which was only meant to be used for test purposes (see section 11 for further details). It comprises 1.927 records. In figure 6, we can also see the link rates for these censuses. Overall, they are substantially lower as those censuses did not have birth place, making much more challenging the unique identification of individuals. Additionally, there are very large gaps between some of the censuses (14 years for the pairs 1787-1801 and 33 years for the pairs 1801-1834)

## 6.2 Rule-based linking in release 1

We developed a relatively simple early rule-based approach to link Danish sources in the early stages of the project (2019-2022). Given the development of machine learning methods afterwards, made possible for our growing set of linked data, we did not update the methods. Thus, the description below is the largely the same one provided in the guide for Link-Lives v.1, released in June 2022, preserved as legacy data.<sup>27</sup>

### 6.2.1 The algorithm

The rule-based links were created using a set of rules determining what constitutes a link. That means that every potential pair of records have been assigned a score which determines the likelihood of the link between them. Following this approach, a threshold value has been set in an attempt to sort *correct* links from *incorrect* links. In this section we present an outline of the algorithm producing the rule-based links, as well as link rates and other performance measures related to the quality of the procedure. The algorithm features 5 major steps:

1. Blocking, i.e. reducing the number of potential links.
2. Calculating the similarity of names and birth place.
3. Calculating a combined link score.
4. Applying a threshold to sort links from non-links.
5. Making household links.

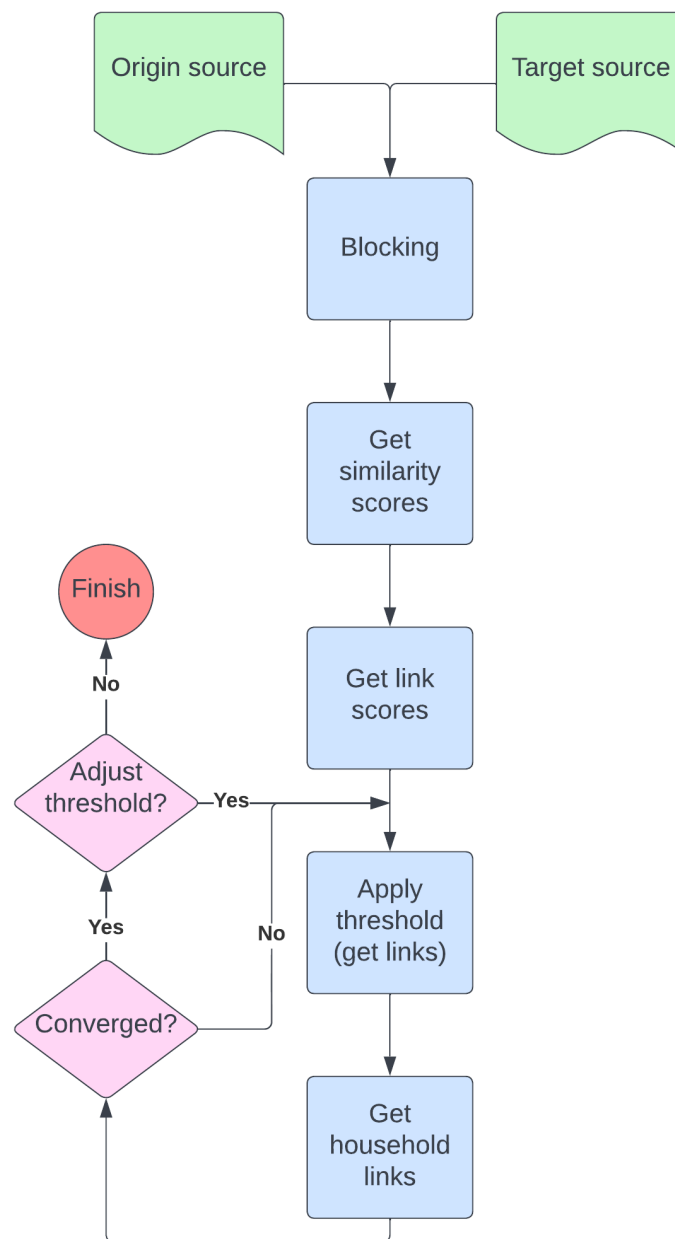
Step 4 and 5 are repeated until no further links are produced by repeating. After reaching this convergence, the threshold parameters are adjusted to allow for more uncertain links to be considered, thus repeating the process until all present thresholds have been used. See figure 7 for a visual representation of the algorithm.

**6.2.1.1 Blocking** The most naive approach to linking would start by making the Cartesian product between the origin source and the target source, i.e. all possible pairs of records, and then proceed to calculate similarity scores. However, given the quadratic scaling, this very quickly becomes unfeasible as the size of the dataset grows. Instead, it is common practice to reduce the number of “potential” links to be considered, when calculating the similarity, by only including potential links that fit certain criteria. This process is known as blocking. We set the criteria that potential matching records must have to the same sex and an age within  $\pm 2$  years to pass the blocking stage. Due to the age criteria, anyone with no age (or birth year) is at not considered as a link candidate. This means that most often, only the “key” person(s) in the parish registers (see ) can be linked.

---

27. The rule-based links from release 1 can be found in the file “links-v1.2.csv”

Figure 7: The rule-based algorithm.



**6.2.1.2 Similarity scores** We use the Jaro-Winkler distance for comparing name and birth place strings. For names, the score is based on comparing the standardised and categorised first names, family names and patronyms. For the birth place, the score is based on the standardised and categorised birth place variables. See section 6.3.1 under **encoding** for more details on how the similarity cores are computed.

**6.2.1.3 Link score** The link score is simply the sum of similarity scores. Since we have used the Jaro-Winkler distance, a score of zero denotes a perfect match.

**6.2.1.4 Applying thresholds (primary links)** After calculating link scores for all potential links passing the blocking stage, we establish primary links. Our initial threshold is a link score of maximum 0.03 and a requirement that there is only one unique link fulfilling this criteria. In subsequent iterations, the tolerance is increased in steps of 0.02 until a maximum of 0.15 is reached. In the censuses, all persons are organized into households. If there are multiple potential links with scores below the tolerance, household information is used to disambiguate the decision.

**6.2.1.5 Household links** The primary links will implicitly also link households. Therefore we allow for links between already linked households to be made with a more relaxed tolerance in a step following the decision on what a primary link is. These are referred to as household links.

**6.2.1.6 Iteration and tolerance adjustment** After having made household links, some of the potential links that were ambiguous before, now possibly stand out as unique links. Therefore we iterate over these two steps until convergence. After reaching convergence in this inner loop, the tolerance is adjusted, as described above, and the process starts again.

## 6.2.2 Link rates

There are currently 5.5 million rule-based links in the database. The distribution can be seen in figure 8 and 9 and the link rates are shown in 10. The figures show that most links are made in the period 1845-1901 (the censuses featuring individuals' birth places) while the early period 1787-1845 is not as well represented. The rule-base approach does not perform well with the parish records and the link rates are under 10% so we do not show specific link rates.

## 6.2.3 Quality testing

We can measure the quality of the rule-based links by comparing them to the bechmmark dataset created manually. In this period, the best result is for the 1850 to 1845 pairing, where the precision is 95-97% with a recall of 53-58%. The worst case is 1901 to 1880 with a precision of 82-85% and a recall of 25-30%.

Figure 8: Total count of links from census to census and from Copenhagen burial register to census.

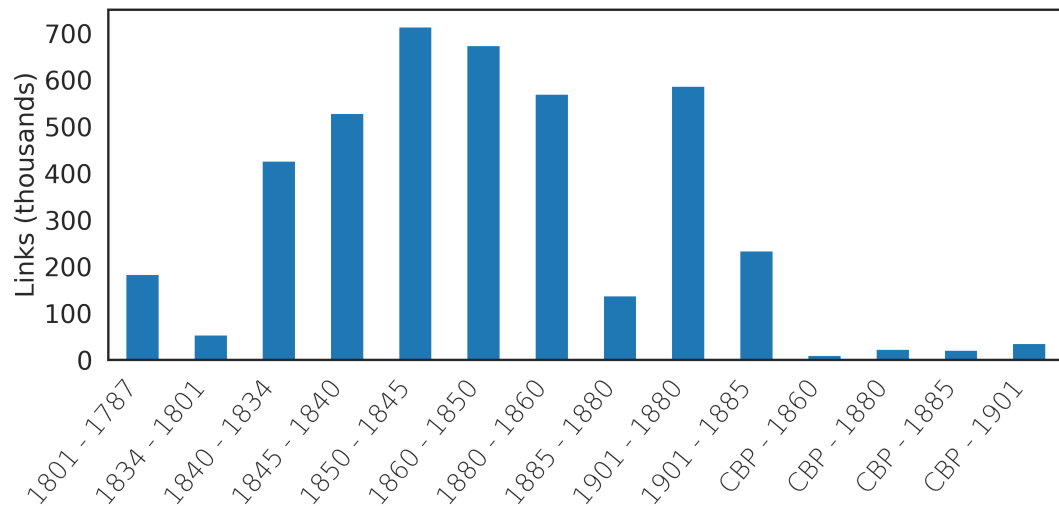


Figure 9: Total count of links from PR to censuses.

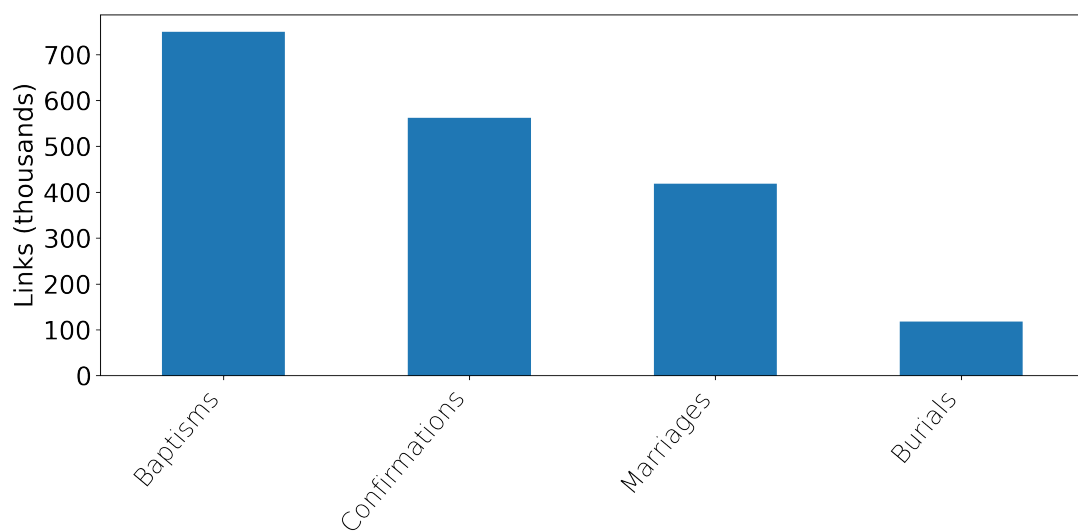
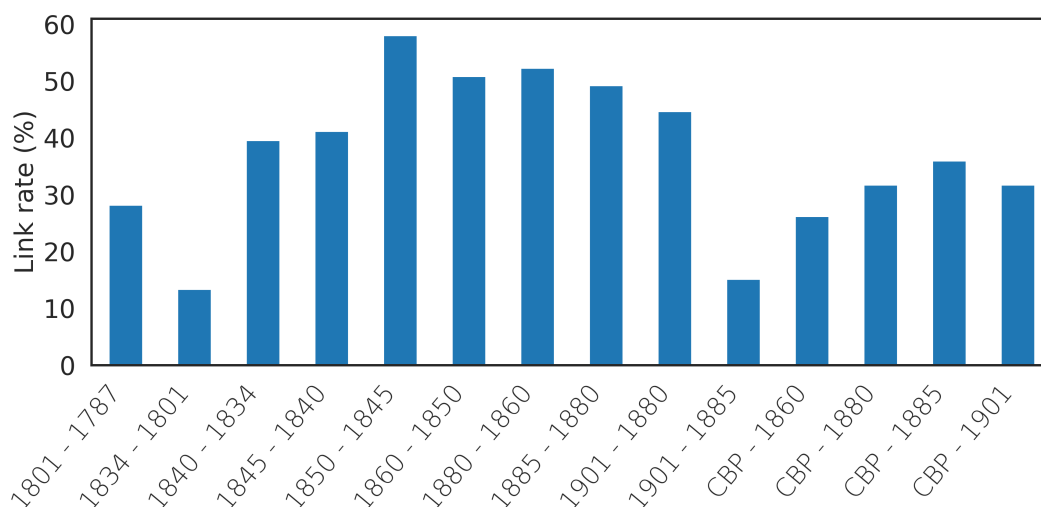


Figure 10: Link rates of links made from censuses and Copenhagen burial register.



### 6.3 Machine learning linking in release 2

While the rule-based linking approach from section 6.2 is explainable and was able to automatically link between many sources, it is ultimately limited in its ability to mimic the approach taken by domain experts during manual linking. As a result, the precision and linking rates of the rule-based approach are sub-optimal compared to those of the manual links. This section introduces a machine learning framework as an alternative method to the rule-based approach, which has the potential to surpass the former approach by learning how to link given a series of examples, known as “supervised learning”<sup>28</sup>. The idea behind supervised learning is to rely on an annotated dataset, in this case, potential pairs of records identified by the domain-experts (or annotated in the machine learning language) as “link” or “not link”. In section 6.1 we describe our annotated dataset, that we call a “benchmark dataset”. The machine learning approach then uses a model that can be trained using the annotated dataset in order to mimic the approach done by the domain experts. The model used here is inspired by an approach used in a Norwegian context Park 2022, that utilized Extreme Gradient Boosting (XGB) model for linking. XGB is a flexible and popular model that relies on decision trees to model complex relations between input and output. It also has the advantage of being fast to train and evaluate compared to other machine learning models such as neural networks. See more about XGB in the original paper (Chen and Guestrin 2016), and see <https://xgboost.readthedocs.io/> for the implementation that is used for linking. Using XGB, our goal is to keep precision over 95% of correct links, to ensure high-quality links. The main steps remain the same as in rule-based linking (comparing origin and target, blocking, getting similarity scores, applying rules), although the last step of applying rules (here, XGB) is more complex. The following section provides a theoretical introduction to supervised learning and XGB. Specific model choices are introduced from section 6.3.2 onward.

28. The machine learning links from release 2 can be found in the file “links\_v2.1.csv”

### 6.3.1 Introduction to linking with supervised learning and XGB

**6.3.1.1 Supervised learning** The problem of linking records from two sources can be framed mathematically. Say that  $x$  and  $y$  are two records describing, e.g., one person in the 1901 census and one person in the 1880 census. The variables  $x$  and  $y$  can be thought of as dictionaries containing  $m$  items on the form **key** : **value**, where **key** denotes what type of data is provided in **value** (the attributes of an individual). For example:

$$x = \{\text{name : Knud Rosendahl,} \\ \text{birth\_year : 1827,} \\ \text{event\_district : odense,} \\ \dots\}, \quad (2)$$

$$y = \{\text{name : Knud Rosendall,} \\ \text{birth\_year : 1827,} \\ \text{event\_district : københavn,} \\ \dots\}, \quad (3)$$

here showing only the first three items of each record. Furthermore, let us denote the *label* (the decision taken by the domain-experts) for two records as the binary variable  $z \in \{0, 1\}$  where 0 corresponds to “not link” and 1 corresponds to “link”. The goal of the linking procedure is then to determine  $z$  given  $x$  and  $y$ . In the supervised learning context, we start by assuming the availability of an annotated dataset with  $n$  examples in total, denoted

$$\begin{aligned} \mathcal{D}^n &= (X^n, Y^n, Z^n), \\ X^n &= (x_1, \dots, x_n), \\ Y^n &= (y_1, \dots, y_n), \\ Z^n &= (z_1, \dots, z_n), \end{aligned}$$

such that  $z_i$  is the label (0 or 1) for the record pair  $(x_i, y_i)$  for  $i = 1, \dots, n$ . In practice, we obtain the annotated dataset from the domain expert links described in section 6.1. In the machine learning framework, we now seek to learn some function  $f$  that can predict the true label given the input, i.e.,

$$\hat{z} = f(x, y),$$

where  $\hat{z}$  is the predicted label based on the input  $(x, y)$ . The function  $f$  is learned based on the dataset  $\mathcal{D}^n$  by minimizing the objective function that measures the error between the predicted and the true label in the training examples. Mathematically, we want to find a function  $f$  that minimizes the *total loss*  $L$  in the training examples given by

$$L = \sum_{i=1}^n \text{logistic}(z_i, \hat{z}_i), \quad (4)$$

where  $\hat{z}_i = f(x_i, y_i)$  is the predicted label for example  $i$ , and  $\text{logistic}(z_i, \hat{z}_i)$  is the binary logistic error measuring the error between the true label  $z_i$  and the

predicted label  $\hat{z}_i$ . We omit the full mathematical expression here, but understand that the error is high when the labels do not match the predictions ( $z_i \neq \hat{z}_i$ ) and low when they match ( $z_i = \hat{z}_i$ ). Combining the loss from each  $i = 1, \dots, n$ , the sum in (4) is then the total training loss with respect to the training data and the model  $f$ . Finding a model  $f$  that minimizes  $L$  then promotes that the examples in the training data are correctly labeled, which in turn promotes that previously unseen links are also labeled correctly (assuming that the dataset  $\mathcal{D}^n$  is sufficiently large and representative). Before discussing how to find the function  $f$ , we introduce an important initial step called encoding.

**6.3.1.2 Encoding** The first step to determine whether  $(x, y)$  is a link is to encode the pair of records into a numerical representation that characterizes the difference (or some other measure useful to compare them) between the records. For example, birth years could be compared using the absolute difference, e.g., given years 1840 and 1844 in  $x$  and  $y$ , respectively, the absolute difference is:

$$|1840 - 1844| = 4,$$

Note that the encoding step was also used to compute similarity scores in the rule-based method in section 6.2.1, with the only difference here being that we encode more information. In particular, we use an encoding of a size  $k$  that varies depending on the available information between two sources. The result of encoding will be a vector  $v = (v_1, v_2, \dots, v_k)$  of size  $k$ , where each element is a feature describing the difference of chosen values in records  $x$  and  $y$ . In the example of comparing  $x$  in (3) and  $y$  in (2), we could choose to encode the difference between **name**, **birth\_year** and **event\_district** as features such that  $k = 3$  giving

$$v = (1, 0, 0), \tag{5}$$

here counting the number of different letters for **name**, the absolute difference for **birth\_year**, and whether or not **event\_district** match for  $x$  and  $y$  encoded as 1 or 0 (in this case 0). Importantly,  $v$  now only contains numbers and can therefore be used as the input to a mathematical linking model. Choosing which features to use for encoding is a critical step in designing the model. Below is a complete list of methods used for the numerical encoding of records:

- **Jaro–Winkler similarity (JW)**: This metric measures the similarity between strings and is often used in the literature for record comparison. The Jaro-Winkler similarity is 1 if two strings match perfectly (e.g., **Knud Rosendahl** and **Knud Rosendahl**) and 0 when the two strings share no similarity (e.g., **Jens** and **Tom**). For the remaining cases, the score is based on how many characters match and how many *transpositions* there are between the strings. For example the Jaro-Winkler similarity between **Christian** and **Kristiansen** is 0.80 while the score between **Christian** and **Jens** is 0.45. Note that the the Jaro-Winkler similarity slightly favours string starting with the same characters, i.e., prefix, where we use a prefix weight of  $p = 0.1$  — we gently refer to Christen 2012, Ch. 5 for more details on the Jaro-Winkler similarity.



- **Absolute difference (diff):** Given two numerical values  $n_x$  and  $n_y$ , the absolute distance is computed as  $|n_x - n_y|$  similar to the example above. Absolute difference is computed the same as the regular difference, but is always positive. For example,  $|5 - 1| = 4$  and  $|1 - 5| = |-4| = 4$ .
- **Self information (si):** Self-information measures how common some variable (typically a string) is among the whole dataset. For example, the name **Jens** has a low self-information because it is very common, while a name like **Napoleon** has a high self-information being less common. To compute the self-information, e.g., for **first\_name**, we start by counting the fraction  $f$  of occurrences of each unique value in the source. We then define  $si = -\ln(f)$ , where  $\ln$  is the natural log. Finally, self-information is normalized to be between 0 and 1. As an example, the most common first name in the 1890 census is **marie** and is given the self-information of 0. On the other hand, **matias karentine** only occurs twice and is given the self-information of 0.99. Now, since we are comparing two records  $x$  and  $y$ , the self-information from both records are added together thus providing a “joint” self-information measure.
- **Distance (dist):** This metric measures the geographical (straight) distance in kilometers applied on keys that describe locations, e.g., birthplace or event place.
- **Same place (same):** Similar to distance, but defined as 1 when the locations are the same and 0 otherwise.

The specific choices encoded features depend on the model — see table 18 for specific details. For a chosen set of  $k$  features, the data  $X^n$  and  $Y^n$  is then encoded as:

$$V^n = (v_1, \dots, v_n), \quad (6)$$

where  $v_i$  of size  $k$  encodes the difference between  $x_i$  and  $y_i$  for  $i = 1, \dots, k$ . Hence, each  $v$  in  $V^n$  is now a vector of numbers that encodes information on the difference between each record pair  $(x, y)$ . This transformation allows for the encoded data to be fed to a link-classification model  $f$  — in our case XGB.

**6.3.1.3 Decision trees and XGB** XGB is a model that uses a *boosted decision tree* that can perform classification based on a given input. Let us start by explaining what a decision tree is.

A decision tree is a simple model that uses a series of Yes/No questions to determine its prediction. The input to the decision tree is an encoded vector  $v$  and the prediction is denoted  $\hat{z} = f(v)$ , where  $f$  is the decision tree. A decision tree to determine if  $v$  should be classified as a link ( $\hat{z} = 1$ ) or not link ( $\hat{z} = 0$ ) is exemplified in figure 11. The examples have three questions relating to **name**, **birth\_year**, and **event\_district** that enable classifying an input  $v$  as “link” or “not link” depending on the Yes/No answers. In practice, the decision tree must also handle when a value is *missing*, for example, if the birth year is missing from

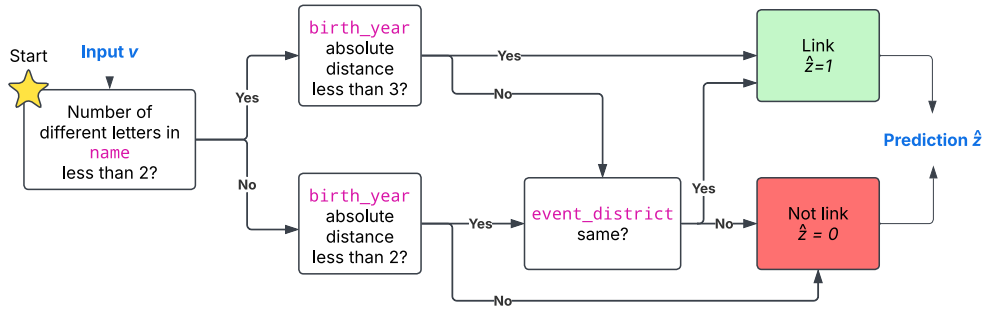


Figure 11: Example of decision three with input  $v$  containing the number of different letters in **name**, the absolute distance of **birth\_year**, and whether or not **event\_district** is the same. Starting from the left, a number of decisions are made based on the features to produce the final prediction  $\hat{z}$

a record. This is handled by adding three branches instead of two for each question, i.e., one branch for answering Yes, No, and Missing to each question. The example in figure 39 is quite simple, but a decision tree can be made arbitrarily large by adding more questions to facilitate increasingly complex decision-making and more input features.

It turns out that constructing a single decision tree that is sufficiently large and complex to correctly predict the labeled training set  $\mathcal{D}^n$  is generally tricky. Another approach is *boosting*, where several small decision trees are constructed to produce a joint output that performs much better than each separate tree. This is done in XGB by producing the trees sequentially, where each new tree is optimized to correct for the error of the previous tree. The idea is the following: The first tree, denoted  $f_1$ , is designed to minimize the loss in (4). Given the loss from the first tree, the second tree  $f_2$  is optimized to minimize the errors produced by  $f_1$ . The two models are now added together to produce a more accurate output, i.e.,  $f(v) = f_1(v) + f_2(v)$ , where the boosted tree  $f$  is the sum of the predictions from  $f_1$  and  $f_2$ . A third tree can now be added to minimize the error produced by the first two. Continuing this strategy, a boosted tree of size  $S$  will make predictions as the sum of each tree

$$f(v) = f_1(v) + f_2(v) + \cdots + f_S(v) = \sum_{s=1}^S f_s(v), \quad (7)$$

where each tree itself can be quite simple (like the one illustrated in figure 11), but the sum of all trees can collectively make very accurate predictions. Note that, while the predictions  $\hat{z}$  should ultimately be 0 or 1, the output of the boosted decision tree is a probability  $p = f(z)$  anywhere between 0 and 1. Here,  $p$  close to 0 indicates “not link” and  $p$  close to 1 indicates “link”, and anything between then indicates the (un)certainty. For example,  $p = 0.8$  indicates that the model predicts a link probability of 80%. We will explain shortly how to go from the probability  $p$  to a prediction  $\hat{z}$ .

The boosted decision tree is constructed based on the encoded training data  $V^n = (v_1, \dots, v_n)$ , the labels  $Z^n = (z_1, \dots, z_n)$ , and a chosen size  $S$  by minimizing the total loss in (4). This is written mathematically as

$$\min_f \sum_{i=1}^n \text{logistic}(z_i, f(v_i)), \quad (8)$$

where minimization is done with respect to the boosted decision tree  $f$  of size  $S$ . Minimization is done based on the mathematical method known as *gradient descent*, where many small changes are made to the model  $f$ , each step lowering the overall loss. Extreme Gradient Boosting (XGB) is an *efficient* implementation of this method for decision trees that uses a few mathematical tricks to enable quick optimization of the model  $f$  on a computer — see Chen and Guestrin 2016 for more details on the underlying algorithm.

**6.3.1.4 Post-calibration of XGB links** So far, we described linking on a pairwise basis, where the link probability  $p$  is evaluated for individual pairs of records from two different sources. This approach works well to build the XGB model, but there are significant gains by considering linking two sources as a whole. For example, when linking two censuses, we generally expect that each person only appears once in each source, and directly enforcing this will help model accuracy. An approach for linking across two sources is simply to link each record in the origin source to the record in the target with the highest predicted score. However, this strategy is generally not advisable since the best candidate can still be a very poor candidate and should be rejected as a link. Additionally, it can be that a record in the origin or target has several good candidates, e.g., with a predicted score over 0.9, which should be carefully resolved.

Hence, assume that two sources  $A$  and  $B$  (say, two censuses) with  $n$  and  $m$  records, respectively, should be linked. Given the XGB model  $f$ , the link probabilities for all possible  $n \times m$  link pairs between the sources could be evaluated. Not all possible pairs will be evaluated due to blocking (as done for the rule-based model), but we shall ignore this aspect for now — see section 6.3.3 for how blocking is done for the machine learning models. We then denote  $p_{i,j}$  as the link probability for linking record  $a_i$  in  $A$  to record  $b_j$  in  $B$  for  $i = 1, \dots, n$  and  $j = 1 \dots, m$ . This is exemplified in figure 12, where  $A$  has  $n = 3$  records and  $B$  has  $m = 4$  records giving a total of 12 possible pairs. Starting from  $a_1$  in the example, we see that it has only one good candidate being  $b_2$  with a link probability of  $p_{1,2} = 0.92 = 92\%$ .  $a_2$  on the other hand, has two good candidates being  $b_1$  with  $p_{2,1} = 0.91$  and  $p_{2,4} = 0.94$ , respectively. Finally,  $a_3$  has a clear best candidate of  $b_4$  with  $p_{3,4} = 0.99$  although it also has  $p_{3,1} = 0.67$  to  $b_1$ . We have implemented the following two rules to deal with linking across two sources:

1. The link score should be the highest value among all possible records in both sources, i.e., a link from  $a_{i^*}$  in source  $A$  to  $b_{j^*}$  in source  $B$  should fulfill:

$$\begin{aligned} p_{i^*,j^*} &> p_{i,j^*} \text{ for } i \neq i^* \text{ (highest value in source A from } b_{j^*}) \\ p_{i^*,j^*} &> p_{i^*,j} \text{ for } j \neq j^* \text{ (highest value in source B from } a_{i^*}) \end{aligned}$$

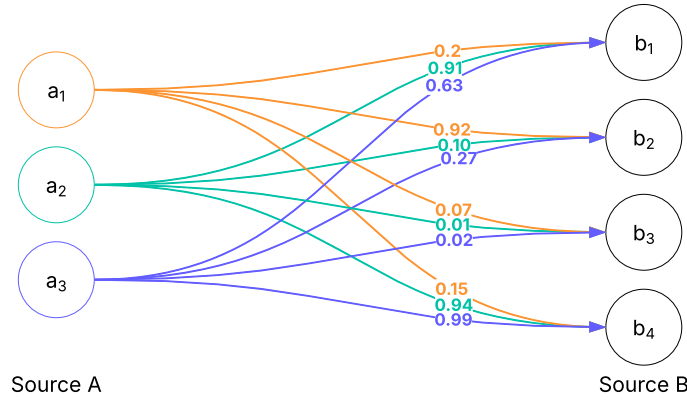


Figure 12: Link probabilities for all record pairs from source A with  $n = 3$  records and source B with  $m = 4$  records. The numbers denote the link probabilities  $p_{i,j}$  for each of the 12 possible pairs.

2. The difference between the highest and second-highest value from source A should be above a threshold  $\delta$ , i.e., if  $a_{i^*}$  in source A has the highest value  $p_{i^*,j^*}$  to  $b_{j^*}$  and second highest value  $p_{i^*,j^{**}}$  to  $b_{j^{**}}$ , then a link should fulfill:

$$p_{i^*,j^*} - p_{i^*,j^{**}} > \delta.$$

Using the first rule ensures that only the best candidate links are used and also prevents linking a record more than once. For example, Hans Kristian Mikkelsen in the 1860 census can be linked to Hans Mikkelsen Sørensen in the 1850 census if no other record from 1860 had a better score to Hans Mikkelsen Sørensen in 1850 and if no other record from 1850 had a higher score to Hans Kristian Mikkelsen Sørensen in 1860. Using the second rule ensures that the link probability is sufficiently high compared to the second-best candidate, where the value  $\delta$  is set manually to achieve the desired balance between precision and recall. In the example of Hans Mikkelsen Sørensen from 1860, we would look at the second match in 1850, for instance, a Mikkelsen Sørensen, and see if the difference in probabilities is larger than the *relative cutoff*  $\delta$ . This means that if two candidates in the target source for a given record are too close to each other, it may be that actually both of them could be the right match, so we do not establish the link. On the contrary, we can capture linked pairs with an overall lower score, that would have never passed an absolute threshold, but that is clearly the best option.

Going back to the example in figure 12, let us assume that  $\delta = 0.3$  has been chosen. We then see that the link from  $a_1$  to  $b_2$  qualifies as a link since:

1.  $p_{1,2} = 0.92$  is the highest value for both  $a_1$  and  $b_2$
2. The highest value from  $a_1$  is  $p_{1,2} = 0.92$  to  $b_2$  and the second highest is  $p_{1,1} = 0.2$  to  $b_1$ , given a difference of  $0.92 - 0.2 = 0.72$ , which is higher than the threshold of 0.3.

For  $a_2$  to  $b_4$  however, we see that both the rules fail since:

1. The highest score for  $a_2$  is  $p_{2,4} = 0.94$  to  $b_4$ , but  $b_4$  has a higher score of  $p_{3,4} = 0.99$  to  $a_3$ .

2. The highest value from  $a_2$  is  $p_{2,4} = 0.94$  to  $b_4$  and the second highest is  $p_{2,1} = 0.91$  to  $b_1$ , given a difference of  $0.94 - 0.91 = 0.03$ , which is lower than the relative cutoff of 0.3.

The link from  $a_3$  to  $b_4$  passes the rules following the same logic (although rule 2 would fail if  $\delta = 0.4$  had been chosen instead). Linking the two sources with these rules then make the link predictions: Predict a link  $\hat{z} = 1$  for  $a_1$  to  $b_2$  and  $a_3$  to  $b_4$  and predict non-links  $\hat{z} = 0$  for all remaining pairs resulting in  $a_2$ ,  $b_1$ , and  $b_3$  not being linked.

The rules for linking one source to another are then applied, and the value of  $\delta$  is calibrated individually for different source pairs. The rules are changed slightly when linking from a source that contains *true duplicates*. For example, when linking marriages to a census, it could be that the same person was married twice. For cases like this, we change rule 1 to:

1. The link score should be the highest value among all possible records in source  $B$ , i.e., a link from  $a_{i^*}$  in source  $A$  to  $b_{j^*}$  in source  $B$  should fulfill:

$$p_{i^*,j^*} > p_{i^*,j} \text{ for } j \neq j^* (\text{highest value to source } B \text{ for } a_{i^*})$$

Rule 2 remains unchanged in these cases. Note that, since we always link to censuses, we never link *to* a source with true duplicates (only *from*).

This concludes the theoretical part of the machine learning models, and the following sections will discuss the specific model designs for linking different sources.

### 6.3.2 Overview of machine learning models used for linking

We created seven models, two of them source-specific (for census-to-census linking), and the remaining five were trained on specific source pairs but designed with the hope that could be re-used for other source pairs. Table 16 shows the models according to their main characteristics and abbreviations. Given the difficulties of linking isolated individuals shown by the rule-based approach, all models (but 1P) attempt to use the presence of multiple individuals in a given event to improve the likelihood of finding the right candidate link: Census-to-census models have features that include information on the co-residents, created by the grouping of individuals in households, without explicitly using family relationships. The other model do not link individuals but pairs and groups of individuals explicitly connected by family relationships (i.e., mother-child, groom-wife).

The models were developed initially for a particular source but then reused for other pairings, as shown in table 17. For instance, the CFM, CF and CM models were developed originally for baptisms and then reused for confirmations and burials. The BG model was developed for marriages. The 1P model was developed for parish records burials and then only re-used on Copenhagen Burials.

We use XGB as the core of each model with different input features and trained with different data. We start by describing the input features used for training the models followed by a description of the data used for training.

Table 16: Overview of models

| Name                            | Abbreviation | Unit of Linking |
|---------------------------------|--------------|-----------------|
| Census with birthplace          | C            | Individual      |
| Census without birthplace       | C0           | Individual      |
| One-person information          | 1P           | Individual      |
| Child-Father-Mother information | CFM          | Group           |
| Child-Mother information        | CM           | Group           |
| Child-Father information        | CF           | Group           |
| Groom-Bride information         | GB           | Group           |

Table 17: Models and source pairs.

| Source pair                     | Models applied  |
|---------------------------------|-----------------|
| Census → Census                 | C               |
| Census → Census (no birthplace) | C0              |
| PR Baptism → Census             | CFM/CM/CF       |
| PR Confirmation → Census        | CFM/CM/CF       |
| PR Marriage → Census            | GB              |
| PR Burial → Census              | CFM/CM/CF/GB/1P |
| PR CBP → Census                 | 1P              |

**6.3.2.1 Features** Comparing records from two different sources requires encoding the information into a number of features as described in section 6.3.1. Each model differs in the chosen features to best utilize the available information. The input features are chosen based on availability, given the limitations on the source itself and in what had been standardized at the time of training. For example, in censuses whose standardization was more complete at the time of the development of the models, we could draw on birth place (harmonized by us) and year, as well as place of residence, which drew on metadata. For parish records, neither **birth\_place** nor **event\_place** had been standardized. Thus, these models are a first step, aimed at testing the feasibility of creating generalizable models based on a limited number of features.

Features are divided into two types:

- About main persons: Features relating to the main person of a given event, according to what each dataset attempts to capture. All persons in a census and Copenhagen burials are considered main persons but only a few of the named persons are main persons in parish records: the child in baptisms,

the married confirmand in confirmations, the bride and groom in marriages and the deceased in burials.

- About support persons: Support persons can be of two types, depending on the sources.<sup>29</sup>
  - For censuses, as every person is a “main person”, we consider other individuals that co-reside in the household as support persons but often we only choose some of them.
  - For other events, we consider the rest of named roles. For instance, father and mother in baptisms, etc.

Table 18 shows the features used as input for the various machine learning models, divided into the variables used for “main persons” and “support persons”. We list the actual name of variable on which we create features, the method used to encode the information and the models where it is used. For instance, we create two features based on the `name_c1` of the “main person”: one computing the Jaro-Winkler distance and one on the self-Information (See section 6.3.1 for an overview of methods used to compute features.) The support persons for which

Table 18: Model features overview

| Type    | Variable                  | Feature     | Model |    |     |    |    |    |    |
|---------|---------------------------|-------------|-------|----|-----|----|----|----|----|
|         |                           |             | C     | C0 | CFM | CM | CF | GB | 1P |
| Main    | <code>name_c1</code>      | JW + SI     | x     | x  | x   | x  | x  | x  | x  |
| -       | <code>first_name</code>   | JW + SI     | x     | x  | x   | x  | x  | x  | x  |
| -       | <code>last_name</code>    | JW + SI     | x     | x  | x   | x  | x  | x  | x  |
| -       | <code>birth_year</code>   | diff        |       |    |     |    |    | x  |    |
| -       | <code>birth_parish</code> | same        | x     |    |     |    |    |    |    |
| -       | <code>birth_town</code>   | same        | x     |    |     |    |    |    |    |
| -       | <code>event_place</code>  | same + dist | x     | x  |     |    |    |    |    |
| Support | <code>name_c1</code>      | JW + SI     | x     | x  | x   | x  | x  | x  |    |
| -       | <code>first_name</code>   | JW + SI     | x     | x  | x   | x  | x  | x  |    |
| -       | <code>last_name</code>    | JW + SI     | x     | x  | x   | x  | x  | x  |    |
| -       | <code>birth_year</code>   | diff        | x     | x  |     |    |    | x  |    |

*JW: Jaro-Winkler; SI: Self-information; diff: difference; dist: distance*

the features are computed are listed in table 19. For example, the CFM model uses father and mother as support persons, each of which is given the features specified in table 18.

Some more details for each model are provided here:

- **CENSUS-TO-CENSUS MODELS:** Census models (C and C0) make a prediction for each individual in the census, using support information of up to four additional persons they co-reside with as supporting information

<sup>29</sup> This division of individuals is close to the “key person” concept introduced in section 6.1.2.3, where “key” individuals are the persons we choose to link manually for our own purposes. However, the difference is that here “main person” refers to the individual on which a record is centered while we use “key” to denote our own choice of linking them or not.

Table 19: Specific support persons used by the models

| <b>Persons</b>  | <b>C</b> | <b>C0</b> | <b>CFM</b> | <b>CM</b> | <b>CF</b> | <b>GB</b> | <b>1P</b> |
|-----------------|----------|-----------|------------|-----------|-----------|-----------|-----------|
| Main person     | All      | All       | Child      | Child     | Child     | Groom     | All       |
| Support Persons |          |           |            |           |           |           |           |
| Nearest above   | x        | x         |            |           |           |           |           |
| Nearest below   | x        | x         |            |           |           |           |           |
| First male      | x        | x         |            |           |           |           |           |
| First female    | x        | x         |            |           |           |           |           |
| Father          |          |           | x          |           | x         |           |           |
| Mother          |          |           | x          | x         |           |           |           |
| Groom/husband   |          |           |            |           |           |           |           |
| Bride/wife      |          |           |            |           |           | x         |           |

(nearest above and nearest below in a household as well as first male and female). This means that the information of a given person can be used by the model multiple times as support information for other household members.

- **CFM, CM, CF:** The CFM model uses available information on the child, the mother and the father from the original source and finds a similar combination of individuals within a target census, so it is a form of group linking. The related CF and CM models were created to account for the fact that families would not necessarily remain cohabitating after the birth of the child (caused by death, migration, work of one of the spouses, etc), so they only look for a combination of children and mothers and children and fathers or that fathers may not necessary be listed in the baptism of the child. They all take the child as the main person, building on the expectation that the child has survived to the next census and can be found with either their father or mother, which was not always the case. An additional mother and father model (a MF model) would be required to complete full linking attempts using multiple information from baptisms to censuses but time constraints prevented us from implementing it for this version. As stated above, they were originally developed with training data from baptisms.
- **GB:** The GB model takes the groom as the main person and only includes information about the wife. It was developed for marriages.

### 6.3.3 Linking direction and blocking

As the models were created to take advantage of a combination of individuals in groups, the linking direction maximized linking rates in two ways: for censuses



and burials, we chose to link backward to ensure that the individuals were still alive. However, for baptisms, confirmations and marriages, we chose to link forward, as it would be only possible to find the family groups after the birth or marriage individuals (child and parents or couples). Confirmations could have actually been linked both forward and backward but we chose forward to comply with the baptisms approach, whose model it used.

In order to reduce the size of the comparison, we implemented a restrictive blocking approach, caused by time and computing restrictions, summarized in table 20. The consequence of this decision is that missing data on three variables prevents many comparisons to be made and reduce the scope of what we can link (given the size of missing data in some of the variables).

Table 20: Summary of blocking parameters and direction.

| Source Pair            | Sex | Birth Year    | Initials | Event Year (Direction) |
|------------------------|-----|---------------|----------|------------------------|
| Census-to-Census       | X   | $\pm 2$ years | X        | Backwards              |
| Baptism-to-Census      | X   | $\pm 2$ years | X        | Forward                |
| Confirmation-to-Census | X   | $\pm 2$ years | X        | Forward                |
| Marriage-to-Census     | X   | $\pm 2$ years |          | Forward                |
| Burial-to-Census       | X   | $\pm 2$ years | X        | Backwards              |
| CBP-to-Census          | X   | $\pm 2$ years | X        | Backwards              |

#### 6.3.4 Training and calibration data

The data for training the models (denoted  $\mathcal{D}^n$  in section 6.3.1) is based on labeled datasets generated using domain expertise. The nuanced linking approach described in section 6.1 is reduced to categorizing record pairs as “link” or “not link”. Table 21 summarizes from which source pairs the training data for each model is generated. We created some of the models with the exact training data required for the intended source pairs but we also experimented with models trained on recycled and derived data to test the limits of generalization of the data we created.

We note that:

Table 21: Training Data Sources for Models.

| Model     | Training Sources  | Data |
|-----------|-------------------|------|
| C/C0      | 1901 → 1880       |      |
|           | 1880 → 1860       |      |
|           | 1860 → 1850       |      |
|           | 1850 → 1845       |      |
| CFM/CM/CF | Baptism → Census  |      |
| GB        | Marriage → Census |      |
| 1P        | Census → Census   |      |
|           | Baptism → Census  |      |
|           | Marriage → Census |      |
|           | Burial → Census   |      |

- **Both census models** were trained by pooling data from four of our census-to-census benchmark dataset for the period 1845-1901. Given that at the time of developing the models, we had not acquired the benchmark dataset for the period 1787-1845, the C0 model was trained to experiment with the possibility of re-using data the same data for C by removing the feature **onbirth place**.
- The **CFM, CF and CM models** were trained using the same data. Given that we did not have sufficient examples of mother-child and father-child only links from the training data, we experimented with training the models on the full data by removing either the father or the mother.
- The **GB model** was trained using data exclusively from marriages but the **1P model** used data from all the other source pairs by focusing only on the main individual.

**6.3.4.1 Generation of additional non-links** Since a positive link between two sources necessarily excludes all other links to the target source, we can use data on links to generate additional non-links for training. Hence, for each link in the dataset, we generate a large number of non-link examples. The non-links are chosen as close matches to the link. For example: Johanne Olsen, born 1830 in Stevnstrup in the original source has been linked to Johanne Olesen, born 1830 in Stevnstrup in the target source. We can then select 100 other persons in the target source that are also named Johanne Olesen (or something similar) and label them as non-link, since these are necessarily not “the right link”<sup>30</sup> in the origin

30. See 6.1 for our brief discussion of human linked data as ground truth

source. Similarly, we can choose 100 persons born in 1830 and 100 persons born in Stevnstrup. By choosing these “close non-links”, we challenge the machine learning model to correctly classify non-links that could look like links. This strategy significantly increases the precision of the model after training. Table 24 summarizes how the non-links are chosen in the different models for each link in the labeled data, where “Type” indicates whether the information refers to a Main or Support person, “ $n$ ” is how many non-links are selected for each link, and “method” indicates what criteria is used to select the best close non-links on the “variables” column (see section 6.3.1 for a description of methods). Note that

Table 22: Selection of non-links for each link in labeled data.

| Model     | Type           | n   | Method   | Variables   |
|-----------|----------------|-----|----------|---|
| C         | Main           | 100 | JW, dist | <b>name_cl</b> ,<br><b>birth_place</b>                      |
| -         | Support        | 20  | JW       | <b>first_name</b> ,<br><b>last_name</b>                     |
| C0        | Main           | 100 | JW       | <b>name_cl</b>  |
| -         | Support        | 20  | JW       | <b>first_name</b> ,<br><b>last_name</b>                     |
| CFM/CM/CF | Main           | 500 | JW       | <b>name_cl</b> ,<br><b>first_name</b> ,<br><b>last_name</b> |
| -         | Support        | 50  | JW       | <b>name_cl</b> ,<br><b>first_name</b> ,<br><b>last_name</b> |
| -         | Main + Support | 50  | JW       | <b>name_cl</b> (main)<br>+ <b>name_cl</b><br>(support)      |
| GB        | Main           | 500 | JW       | <b>name_cl</b>  |
| -         | Support        | 50  | JW       | <b>name_cl</b>  |
| 1P        | Main           | 500 | JW       | <b>name_cl</b>  |

number of non-links for the support persons provided in the table is created for each support person specified in table 19. For example, for model C, 20 non-links are chosen according to the minimum Jaro-Winkler (JW) distance for **first\_name** and **last\_name** for each of the support persons, which are *nearest above*, *nearest below*, **first male**, and **first female**. The non-links for the CFM/CM/CF models where it lists “Main + Support” should be understood as the 50 closest matches in JW distance for both the **name\_cl** of the main person and **name\_cl** of the support person.

These non-links are not sampled from the entire target source but within blocks. However, the blocking strategy for generating training data is slightly different than the one used for full-scale linking, as 23 shows. We remove initials and increase birth year range for some models (see table 20 for choices in full linking).

Table 23: Blocking Approach for Different Models for Training Data Construction.

| Model     | Sex | Birth Year    | Event Year (Direction) |
|-----------|-----|---------------|------------------------|
| C         | X   | $\pm 2$ years |                        |
| C0        | X   | $\pm 2$ years |                        |
| CFM/CM/CF | X   | $\pm 3$ years | Forward                |
| GB        | X   | $\pm 3$ years | Forward                |
| 1P        | X   | $\pm 3$ years | None / Forward         |

The strategy of using domain expert links and close non-links allows us to create a large training dataset that approximates the distribution between links and non-links. Table 24 gives an example of the size of the training data that was used to train the C and C0 models.

Table 24: Model C and C0 sample of links and non-links.

| Source pair             | Links | Non-links | Total     |
|-------------------------|-------|-----------|-----------|
| 1901 $\rightarrow$ 1880 | 682   | 400,599   | 401,281   |
| 1880 $\rightarrow$ 1860 | 712   | 413,764   | 414,476   |
| 1860 $\rightarrow$ 1850 | 791   | 403,660   | 404,451   |
| 1850 $\rightarrow$ 1845 | 855   | 408,064   | 408,919   |
| <b>Total</b>            | 3,040 | 1,626,087 | 1,629,127 |

After generating the training data, we split it into two parts: “model training data” and “calibration data”. The model training data is used to train the XGB models, while the calibration data is used to calibrate block linking (i.e., selecting the threshold  $\delta$  in the post-calibration). We generally dedicated 50% for model training data and 50% for calibration data.

### 6.3.5 Selecting threshold $\delta$ for post-calibration of XGB links

With the XGB models fitted and evaluated for all record pairs within the blocks, the last step is to classify links and non-links according to the rules described in the post-calibration step described in section 6.3.1. In particular, the threshold  $\delta$  for the minimum score difference between best and next best candidate should be chosen for each model. The threshold is chosen by evaluating the precision and recall on the calibration data for different thresholds. We aimed at selecting a  $\delta$  that achieves a precision of 95% on the calibration data. Note that the precision achieved on the calibration data might not be representative of the true precision

due to the dependency introduced by systematically exploring for the threshold. We therefore also perform extensive validation on the generated links to verify the precision (see section 6.3.7). The results of the calibration are presented below.

**6.3.5.1 Census models C and C0** We established the relative cut-off  $\delta$  for censuses for the four test sets we extracted from our own benchmark dataset to reach our target of 95% precision. For the remaining census pairings, we examined the results of the model results visually and applied thresholds that yielded similar results. Table 25 summarizes the selected values of  $\delta$ , achieved precision, and recall on the calibration data.

Table 25: Precision and Recall for Different Source Pairs Evaluated on Calibration Data.

| Source Pair             | Relative Cutoff $\delta$ | Precision (%) | Recall (%) |
|-------------------------|--------------------------|---------------|------------|
| 1901 $\rightarrow$ 1880 | 0.6                      | 95.6          | 36.6       |
| 1901 $\rightarrow$ 1885 | 0.5                      | -             | -          |
| 1885 $\rightarrow$ 1880 | 0.5                      | -             | -          |
| 1880 $\rightarrow$ 1860 | 0.7                      | 95.1          | 34.8       |
| 1860 $\rightarrow$ 1850 | 0.1                      | 95.1          | 75.7       |
| 1850 $\rightarrow$ 1845 | 0.1                      | 95.6          | 76.7       |
| 1845 $\rightarrow$ 1840 | 0.2                      | -             | -          |
| 1840 $\rightarrow$ 1834 | 0.2                      | -             | -          |
| 1834 $\rightarrow$ 1801 | 0.5                      | -             | -          |
| 1801 $\rightarrow$ 1787 | 0.5                      | -             | -          |

**6.3.5.2 Parish records and Copenhagen burials models** For parish records and Copenhagen burials, we followed a similar approach and we allowed a more stable threshold across all models (see table 26). However, many of these models were calibrated with relatively small datasets, as shown in the last column. As there were no calibration data for confirmations, the threshold were chosen based on visual inspection of the data to align with the threshold set for the baptism to census pairings.

**6.3.5.3 Resolving link conflicts due to multiple models** As seen in table 17 and 26, some source pairs have been linked using more than one model. For example, baptisms to census have been linked by both the CFM, CM, and CF models to possibly link events where either the father or mother is missing. This naturally leads to inconsistencies, which arise when two models proposing different candidates in the target source for a given person-record in the origin

Table 26: Precision, Recall, and Test Links from Parish Records to Censuses evaluated on Calibration data

| Source Pair (Model)                           | Relative Cutoff $\delta$ | Precision (%) | Recall (%) | Number of Calibration Links |
|---|--------------------------|---------------|------------|-----------------------------|
| Baptism $\rightarrow$ Census (CFM)            | 0.8                      | 95.5          | 32.3       | 393                         |
| Baptism $\rightarrow$ Census (CM)             | 0.8                      | 95.3          | 20.9       | 465                         |
| Baptism $\rightarrow$ Census (CF)             | 0.8                      | 89.2          | 17.5       | 439                         |
| Confirmation $\rightarrow$ Census (CFM/CM/CF) | 0.8                      | -             | -          | 0                           |
| Marriage $\rightarrow$ Census (GB)            | 0.2                      | 96.6          | 49.3       | 292                         |
| Burial $\rightarrow$ Census (CFM)             | 0.7                      | 97.6          | 64.6       | 65                          |
| Burial $\rightarrow$ Census (CM)              | 0.7                      | 97.0          | 33.3       | 90                          |
| Burial $\rightarrow$ Census (CF)              | 0.7                      | 96.6          | 24.0       | 395                         |
| Burial $\rightarrow$ Census (GB)              | 0.7                      | 94.6          | 46.8       | 94                          |
| Burial $\rightarrow$ Census (BG)              | 0.7                      | 95.8          | 40.9       | 408                         |
| Burial $\rightarrow$ Census (1P)              | 0.7                      | 94.9          | 9.7        | 1386                        |
| CBP $\rightarrow$ Census (1P)                 | 0.7                      | 95.0          | 30.4       | 1516                        |

source, e.g., the CFM and the CF models link a child in a baptism to different groups of persons in a census. We resolve these inconsistencies differently for the source pairs:

- **PR Baptism  $\rightarrow$  Census (CFM/CM/CF) and PR Confirmation  $\rightarrow$  Census (CFM/CM/CF):** The results of the CFM model are preferred over those of any of the other models, even if they are inconsistent with the results of CF and CM, as they utilize the most information about child, father and mother. CF and CM model results are discarded if they are inconsistent with each other, that is, a separate link to census is found for both mother + child and father + child.
- **PR Burial  $\rightarrow$  Census (CFM/CM/CF/GB/1P):** For burials, links are automatically discarded if there are inconsistencies from the results of different models.

### 6.3.6 Results: Link rates

**6.3.6.1 Overall link rates - per model per year range** Aggregated measures of link rate performance hide the high degree of heterogeneity of the linking approach. We have five events (censuses, baptisms, marriages, confirmations, and burials), a long chronology (1787-1917), different spacing between the censuses, and seven models. To provide an overview, we present a series of heatmaps in the

following table 27 and figures 14 to 25, displaying the link rates for the different models within the five events for the full chronology. To simplify the representation, we have presented link rates broken down in a chronology marked by the census years. As census years are not evenly distributed over our period, the link rates are not directly comparable as the link rate between 1845-1850 will be always lower than the link rate for the period 1860-1880. It is well-known that the more time that elapses between two records the less likely is the possibility of finding a link. And those results are also seen in our figures: overall, we find lower link rates between censuses when we have larger intervals between them (as in 1860-1880 and 1880-1901). That effect is also in place when linking parish records and censuses, as link rates are higher in general to the closest census to an event (but for the case of confirmations) when we find them sometimes two census afterwards. Additionally, we can see that the increase on the amount of information recorded in dataset for individuals after 1892 also improves link rates for the later period. Link rates are really low for burials in general, given the fact that we only have information on one individual and the fact that we do not use place of residence. The exception is the CMF model after 1892 that is able to leverage information on parents and capture links in a census decades before the death.

Table 27: Link Rates and Number of Linkable Records between Consecutive Censuses

| Census Pair | Link Rate (%) | Number of Linkable Records |
|-------------|---------------|----------------------------|
| 1801 → 1787 | 20.28         | 937,164                    |
| 1834 → 1801 | 8.42          | 1,137,729                  |
| 1840 → 1834 | 45.40         | 1,265,193                  |
| 1845 → 1840 | 45.81         | 1,465,862                  |
| 1850 → 1845 | 59.75         | 1,404,046                  |
| 1860 → 1850 | 43.55         | 1,746,767                  |
| 1880 → 1860 | 15.94         | 1,920,753                  |
| 1901 → 1880 | 14.74         | 2,369,222                  |







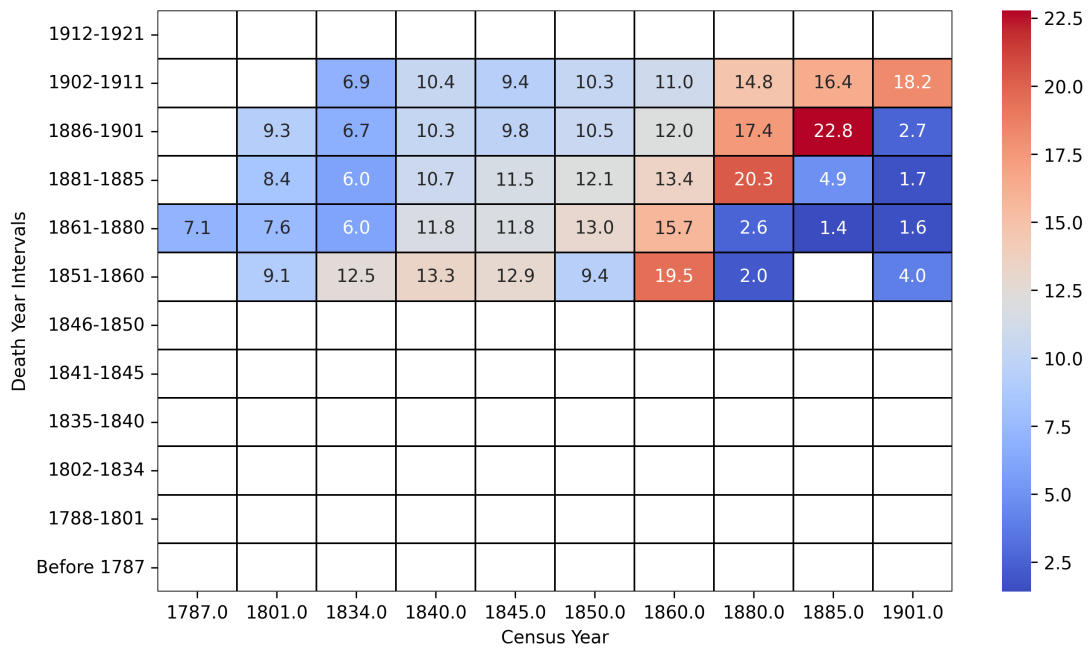








Figure 25: Link Rates for Burials in Copenhagen Burials 1P model



### 6.3.6.2 Representativity assessment

We have calculated link rates by gender and age for each type of source pairings to give a non-exhaustive introduction of the the representativity of the models according to these two main deomographic variables. Figure 26 shows the results for censuses and figure 27 for all parish records and Copenhagen burials where bars below one represent cases where the link rates for women are lower than those for me. The census models shows the expected preference for men that has been shown elsewhere in other contexts, but the CFM model tends to slightly favour women in baptisms and to some extent in confirmations but not burials. CM and CF do not behave systematically in either baptisms or confirmations and 1P tends to prioritize also men but for some outliers in pre-1787 outliers.

Figure 26: Ratio of female to male link rates for different census pairings

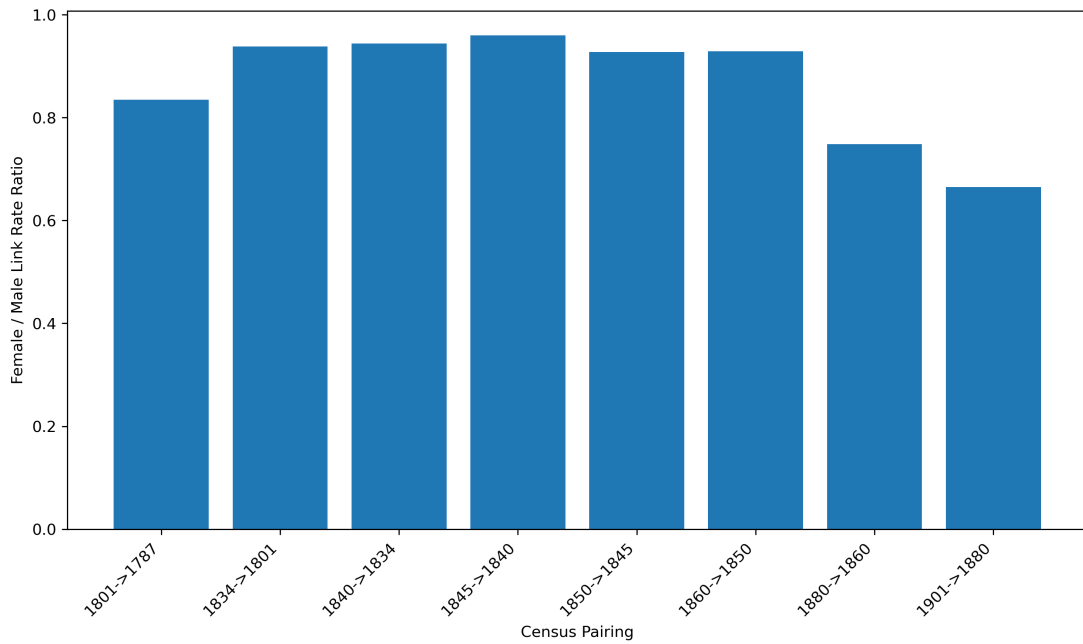
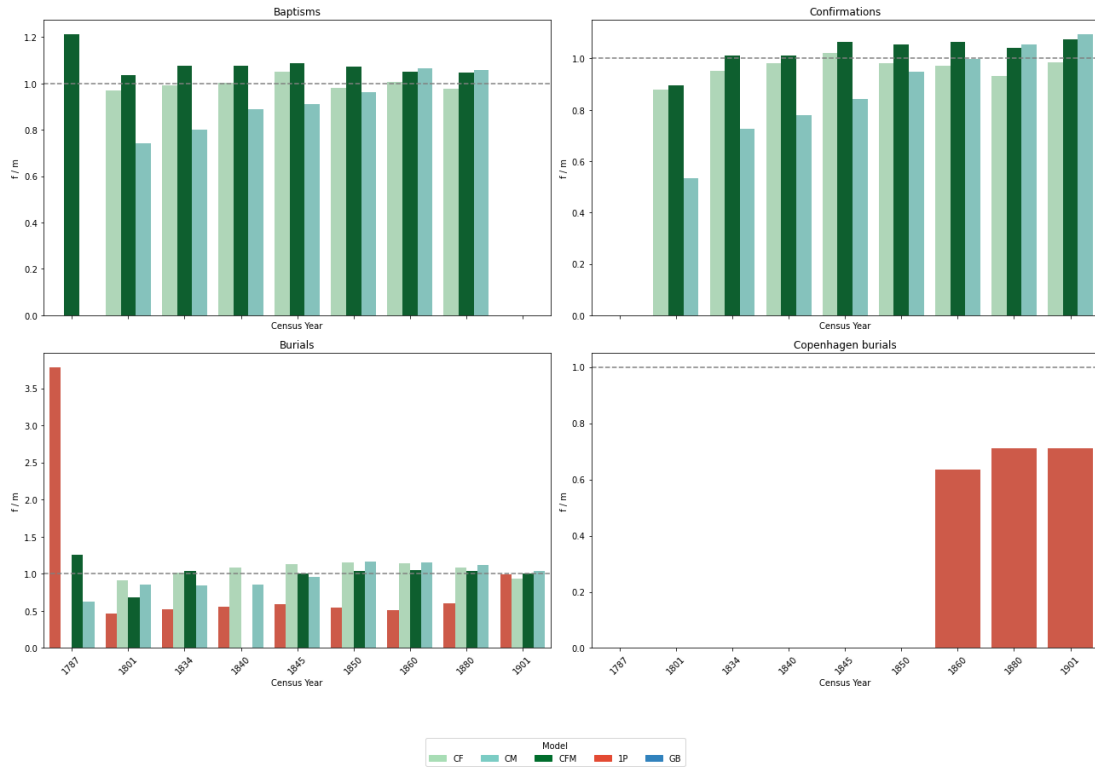


Figure 27: Ratio of female to male link rates for different events, models and census years



*Note: results from model GB for burials have been ommitted for display purposes. The GB model disproportionally favours males with ratios up to 30*

Regarding age, figure 28 shows the results for censuses. Overall, censuses show reduced link rates for individuals in the young adult ages (around their 20s-30s) consistent with the pattern of leaving home and finding employment/getting married somewhere else, which places them in two different household arrangements in two separate censuses and thus, make them harder to be identified. Figure 29 on confirmations shows an inverted U shape in confirmations all over the chronology that actually responds to the same pattern. Ages around confirmation are the latest time a child can be found in the parental home co-residing with both parents. We only display the CFM model but CF and CM models performed similarly. Marriages (not shown) show higher link rates around the 20s-30s, close to the marriage date and with both members of the couple alive. For burials, the link rates over ages vary a lot depending on the model and the year they link to, as it can be seen in figure 30. 1787 has been omitted as only 1P was in use and only linked individuals between 11 and 30. Copenhagen burials results are shown in figure 31.

Figure 28: Link rates by age for different census pairings

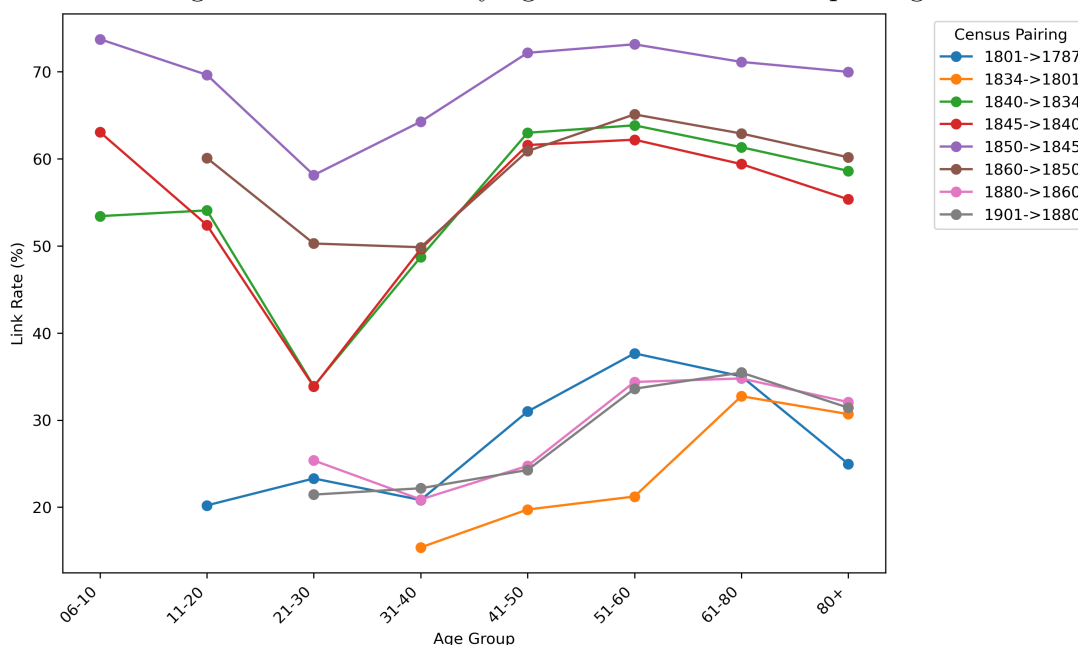




Figure 29: Link rates by age for confirmations for all target censuses (model CFM)

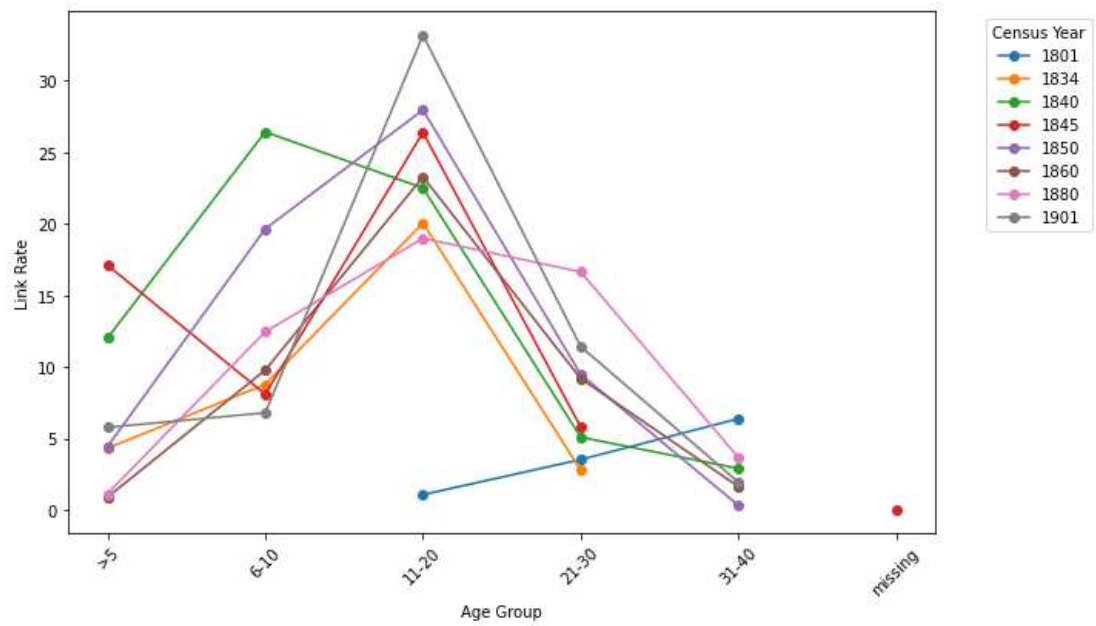


Figure 30: Link rates by age for burials for target census 1801-1901 for all models)

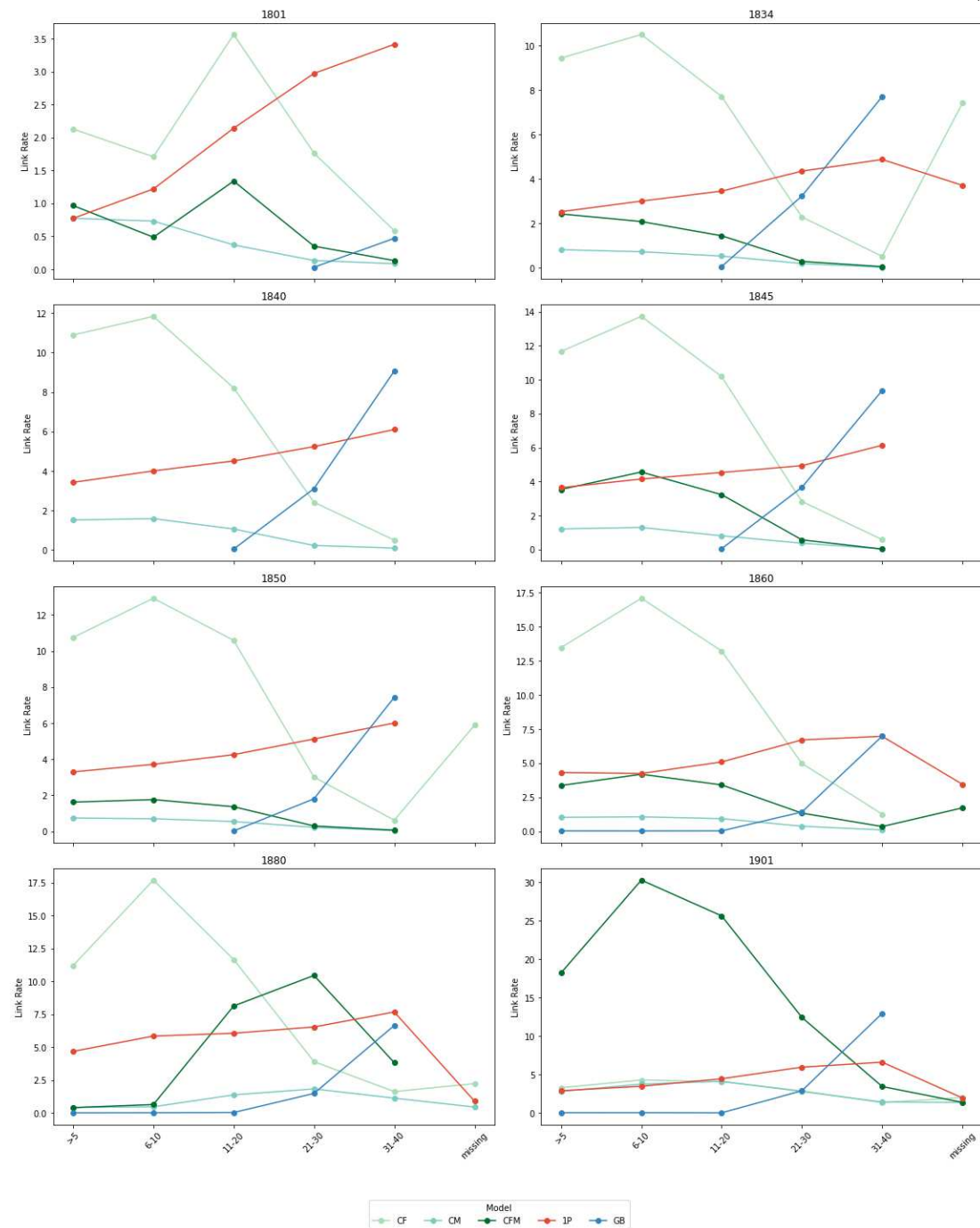
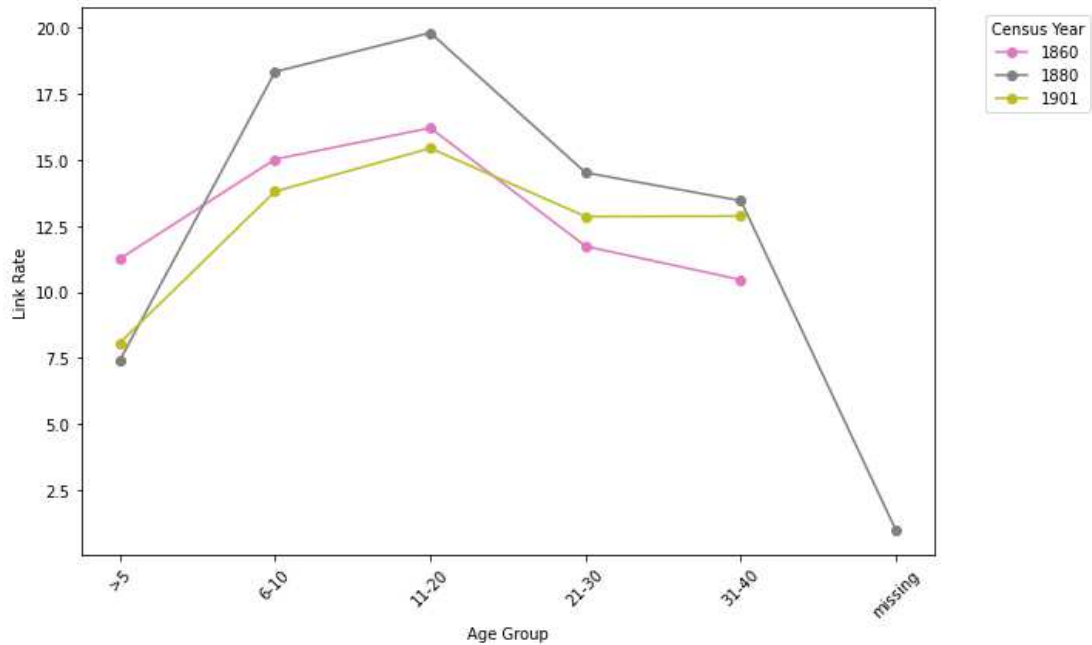


Figure 31: Link rates by age for burials in Copenhagen for target census 1860-1901 for model 1P



### 6.3.7 Validation

We carried out an independent validation of the results for three reasons.

- We have trained three sets of models (C0, CF and CM) on a subset of the variables used for linking to mimic the data omissions in some specific source pairings where we did not have training data. That meant that model estimates did not really reflect the original data. As described above, this affected:
  - The model C0 for the censuses 1787-1845 used training data for the post-1845 period removing the variable birthplace from training.
  - The CF and CM models for births used only data on child and mother or child and father even though the links may have been created using information from all three.
- We used our data partially for some part of the calibration of the models. Given the small amount of data we had for some of the models, unforeseen selection and representation issues in the training data may make the models perform poorly in the full dataset.
- We lacked data entirely for the test of the censuses 1787-1845 and confirmations.

Thus, we considered advisable to carry out an additional manual validation of random samples to assess quality beyond the constraints of our data and ensure

that we only kept those models with higher precision. This was especially important, given that links are later aggregated in lifecourses and errors accumulate and may create inconsistent lifecourses that risk to be entirely omitted.

### 6.3.7.1 Sampling

We designed a sampling strategy that could address some of the most difficult challenges of assessing the quality of five events, seven models and a very long chronology (1787-1917) with changes in available information on the source and transcription quality. A simple random validation of a few hundred cases would not be sufficient to understand the validity of the models and would mix all types of phenomena.

A thorough assessment would have required at least 100 cases for each model within each pairing (original source to each target census, for instance, parish records to census 1787), which would have meant validating more than 46,000 linked records. As we did not have sufficient resources to do that, we chose three specific chronologies systematically (representative of different source, data and societal conditions) with only one linking direction (forwards or backwards). However, instead of samples from that chronology to all possible censuses, we started validating samples from those we assessed to be more likely reliable (based on our knowledge of the data) and we progressively extended to those where we expected worse results. So when the results showed systematically that models were not performing well in general, we refrained from assessing them further in all other potential settings.

For instance, a birth in 1848 could be theoretically linked to all the following censuses but we chose to validate births from 1848-1850 first to the next census, 1850. We only proceeded to validate that cohort of births to 1860 if the results of the links to 1850 indicated that the links were of sufficient quality.

In taking this decision, we also considered whether the model's contribution to the overall link rates for the given event type and time period were sufficient for us to undertake validation (this is for example the case with the Child-Mother model for births and burials and the Child-Father model for births and confirmations). We focused our validation efforts on the models we judged to hold the most potential, both in terms of link rates and link quality. This also means that we validated the models that we eventually kept more thoroughly than the models we dismissed.

For parish records, the two decades we chose to validate reflect the difference in linking conditions between the first and second half of the 19th century. For marriages, baptisms, and confirmations, we chose 1850-1860 and 1824-34, where the nearest target censuses were 1860 and 1834 respectively. For the burials, since they were linked backwards instead of forwards, we chose to keep the 1860 and 1834 censuses as the reference points, and selected burial events from 1860-70 and 1834-44 respectively.

For the burials specifically, we also sampled events from a decade at the very end of the period (1901-1911), due to a change in registration practice in the 1890s, which meant that burial registrations consistently included the names of both mother, father, and spouse (where relevant), regardless of the deceased's

gender or age at death.

Table 28 contains an overview of the validated samples, with the number of links validated in each sample.

Table 28: Number of links validated in each sample

| event type       | model | event years | 1787 | 1801 | 1834 | 1840 | 1845 | 1850 | 1860 | 1880 | 1885 | 1901 | total        |
|------------------|-------|-------------|------|------|------|------|------|------|------|------|------|------|--------------|
| Census-census    | -     | -           | 98   | 99   | 100  | 100  | 100  | 100  | 100  | 200  | 100  | -    | 997          |
| PR baptisms      | CFM   | 1850-60     | -    | -    | -    | -    | -    | -    | 297  | 300  | -    | -    | 597          |
|                  |       | 1824-34     | -    | -    | 300  | 300  | 300  | 297  | 300  | 204  | -    | -    | 1701         |
|                  | CF    | 1850-60     | -    | -    | -    | -    | -    | -    | 200  | 199  | -    | -    | 399          |
|                  |       | 1824-34     | -    | -    | -    | -    | -    | -    | -    | -    | -    | -    | 0            |
|                  | CM    | 1850-60     | -    | -    | -    | -    | -    | -    | 200  | 200  | -    | -    | 400          |
|                  |       | 1824-34     | -    | -    | 200  | -    | -    | -    | -    | -    | -    | -    | 200          |
| PR marriages     | GB    | 1850-60     | -    | -    | -    | -    | -    | -    | 199  | 200  | -    | 200  | 599          |
|                  |       | 1824-34     | -    | -    | 200  | 200  | 200  | 200  | 200  | -    | -    | -    | 1000         |
| PR confirmations | CFM   | 1850-60     | -    | -    | -    | -    | 300  | 300  | 300  | 300  | -    | 48   | 1248         |
|                  |       | 1824-34     | -    | -    | 300  | 300  | 300  | 300  | 300  | 102  | -    | -    | 1602         |
|                  | CF    | 1850-60     | -    | -    | -    | -    | 198  | 198  | -    | -    | -    | -    | 396          |
|                  |       | 1824-34     | -    | -    | -    | -    | -    | -    | -    | -    | -    | -    | 0            |
|                  | CM    | 1850-60     | -    | -    | -    | -    | 200  | 200  | 200  | 200  | -    | 118  | 918          |
|                  |       | 1824-34     | -    | -    | -    | -    | -    | -    | -    | -    | -    | -    | 0            |
| PR burials       | 1P    | 1901-11     | -    | -    | -    | -    | -    | -    | 100  | 100  | -    | 100  | 300          |
|                  |       | 1860-70     | -    | -    | -    | -    | -    | 100  | 100  | -    | -    | -    | 200          |
|                  |       | 1834-44     | -    | 100  | 100  | -    | -    | -    | -    | -    | -    | -    | 200          |
|                  | GB    | 1901-11     | -    | -    | -    | -    | 200  | 200  | 200  | 200  | -    | 200  | 1000         |
|                  |       | 1860-70     | -    | -    | -    | -    | -    | 200  | 200  | -    | -    | -    | 400          |
|                  |       | 1834-44     | -    | 200  | 200  | -    | -    | -    | -    | -    | -    | -    | 400          |
|                  | CFM   | 1901-11     | -    | -    | 300  | -    | 300  | 300  | 300  | 300  | -    | 299  | 1799         |
|                  |       | 1860-70     | -    | -    | -    | -    | -    | -    | 201  | -    | -    | -    | 201          |
|                  |       | 1834-44     | -    | -    | -    | -    | -    | -    | -    | -    | -    | -    | 0            |
|                  | CF    | 1860-70     | -    | -    | -    | -    | -    | 200  | 200  | -    | -    | -    | 400          |
|                  |       | 1834-44     | -    | 200  | 200  | -    | -    | -    | -    | -    | -    | -    | 400          |
|                  | CM    | 1860-70     | -    | -    | -    | -    | -    | -    | -    | -    | -    | -    | 0            |
|                  |       | 1834-44     | -    | -    | -    | -    | -    | -    | -    | -    | -    | -    | 0            |
| Cph burials      | 1P    | 1901-1911   | -    | -    | -    | -    | -    | -    | -    | 100  | 100  | 98   | 298          |
|                  |       | 1860-70     | -    | -    | -    | -    | -    | -    | -    | -    | -    | -    | 0            |
| total            | -     | -           | 98   | 599  | 1900 | 900  | 2098 | 2595 | 3597 | 2605 | 200  | 1063 | <b>15655</b> |

### 6.3.7.2 Validation guidelines

We used our own developed software ALA (see section 38) to inspect the links created by the different models and validated them manually. We designed a limited domain-expert approach consisting of only one domain expert (and 90% of all validation was done by one team member). We allowed only three options for validation:

- “confirmed link”
- “plausible but other candidates exist”

- “information does not match/link is clearly wrong”

For the censuses, only the sampled link decisions were evaluated, but additional household members were shown in the linking interface as context to help evaluate the plausibility of the sampled machine learning links. In parish records, the samples were larger because we included all the individuals linked by a model, as they were linked in groups (in the CFM model, all three people were linked at once, in CF only child and father, etc.) so all of them were shown in the linking interface as context to help evaluate the plausibility of the sampled links.

### 6.3.7.3 Results

Table 29, 30 and 31 show the results of the manual validation, expressed as precision scores (the percentage of links the domain-expert judged to be correct, a confirmed link) for each sample. As is apparent from the 95% confidence intervals (added for each sample in parenthesis), there is some uncertainty attached to the exact precision scores, given that each sample only contain 100-300 linked individuals.

Table 29: Validation results for census→census links, with 95% confidence intervals

| Census pair | No. links<br>validated | Precision          |
|-------------|------------------------|--------------------|
| 1801 → 1787 | 98                     | 70.4 (60.3 , 79.2) |
| 1834 → 1801 | 99                     | 51.5 (41.3 , 61.7) |
| 1840 → 1834 | 100                    | 85 (76.5 , 91.4)   |
| 1845 → 1840 | 100                    | 88 (80 , 93.6)     |
| 1850 → 1845 | 100                    | 96 (90.1 , 98.9)   |
| 1860 → 1850 | 100                    | 95 (88.2 , 98.4)   |
| 1880 → 1860 | 100                    | 97 (91.5 , 98.6)   |
| 1885 → 1880 | 100                    | 97 (91.5 , 99.4)   |
| 1901 → 1880 | 100                    | 96 (90.1 , 98.9)   |
| 1901 → 1885 | 100                    | 89 (81.2 , 94.4)   |

During the manual validation process for the parish records, we observed a common type of error: the main person had been linked “correctly”<sup>31</sup>, but one or more of the related persons from the event had been incorrectly identified in the census household, usually due to unusual (and often un-categorised) household position strings. To evaluate the impact of this error type, we calculated an additional precision score only for the main persons. The precision score is generally slightly higher than the overall precision score for the sample but not game-changing.

31. See our discussion on human linking as benchmarking in section 6.1

Table 30: Example validation results by target census

| Target census | No. links<br>validated | Precision         |
|---------------|------------------------|-------------------|
| 1880          | 100                    | 90 (82.32, 95,10) |
| 1885          | 100                    | 91 (83.60, 95,80) |
| 1901          | 98                     | 89 (81.17, 94.32) |

The results indicated that linking across very long time spans often led to lower precision and they confirmed our suspicion that the experimental use of subsets of variables in training was sub-optimal and created too many false positives. Unfortunately, it also showed that the limited data we had available was not sufficient for testing some of the models, and the results for deaths were substantially worse than the model estimations, the approach of developing generic models was less successful than the developing results had suggested.

However, it is important to note, in contrast, that the results for the CFM model confirmations (trained exclusively on baptism data) on were really positive.

These results made us consider whether it would be be advisable to include all of these models in later lifecourse reconstruction, which we discuss further in section 7.

Table 31: Precision scores in validated samples for parish registers (correct links as % of all validated links in sample)

| Event type        | Model type | Event years | 1787 | 1801 | 1834              | 1840             | 1845             | 1850             | 1860             | 1880             | 1885 | 1901             |
|-------------------|------------|-------------|------|------|-------------------|------------------|------------------|------------------|------------------|------------------|------|------------------|
| PR baptisms       | CFM        | 1850-60     | -    | -    | -                 | -                | -                | -                | 97.3 (94.8,98.8) | 94.3 (91.1,96.7) | -    |                  |
|                   |            | 1824-34     | -    | -    | 98 (95.7,99.2)    | 99 (97.1,99.8)   | 91.7 (87.9,94.5) | 89.2 (85.1,92.5) | 89.3 (85.3,92.6) | 87.3 (81.9,91.5) |      |                  |
|                   | CF         | 1850-60     | -    | -    | -                 | -                | -                | -                | 81.5 (75.4,86.6) | 48.7 (41.6,55.9) |      |                  |
|                   |            | 1824-34     | -    | -    |                   |                  |                  |                  |                  |                  |      |                  |
|                   | CM         | 1850-60     | -    | -    | -                 | -                | -                | -                | 91.5 (86.7,95.1) | 79.5 (73.2,84.9) | -    |                  |
|                   |            | 1824-34     | -    | -    | 88.5 (83.3, 92.6) |                  |                  |                  |                  |                  |      |                  |
| PR marriages      | GB         | 1850-60     | -    | -    | -                 | -                | -                | -                | 96.5 (92.9,98.6) | 93 (88.5, 96.1)  | -    | 82.5 (76.5,87.5) |
|                   |            | 1824-34     | -    | -    | 94 (89.8,96.9)    | 90 (85.1, 93.8)  | 87 (81.5,91.3)   | 85 (79.3,89.7)   | 86 (80.4,90.5)   |                  |      |                  |
| PR conf-irmations | CFM        | 1850-60     | -    | -    |                   |                  | 98 (95.7, 99.3)  | 98 (95.7,99.3)   | 93.3 (89.9,95.9) | 94 (90.7,96.4)   | -    | 85.4+            |
|                   |            | 1824-34     | -    | -    | 94.3 (91.1,96.6)  | 85.3 (80.8,89.1) | 77.7 (72.5,82.3) | 84.7 (80.1,88.6) | 81.7 (76.8,85.9) | 99* +            |      |                  |
|                   | CF         | 1850-60     | -    | -    |                   |                  | 88.4 (83.1,92.5) | 83.9 (78,88.7)   |                  |                  |      |                  |
|                   |            | 1824-34     | -    | -    |                   |                  |                  |                  |                  |                  |      |                  |
|                   | CM         | 1850-60     | -    | -    |                   |                  |                  | 95.5 (91.6,97.9) | 68 (61.1,74.4)   | 82 (76.0,87.1)   |      | 87.3 (79.9,92.7) |
|                   |            | 1824-34     | -    | -    |                   |                  |                  |                  |                  |                  |      |                  |
| PR burials        | 1P         | 1901-11     | -    | -    |                   |                  |                  |                  | 90 (82.4,95.1)   | 86 (77.6,92.1)   |      | 91 (83.6,95.8)   |



Table 31 continued: Precision scores in validated samples for parish registers.

| Event type | ML type | Event years | 1787           | 1801             | 1834 | 1840             | 1845             | 1850             | 1860               | 1880                | 1885 | 1901             |
|------------|---------|-------------|----------------|------------------|------|------------------|------------------|------------------|--------------------|---------------------|------|------------------|
|            |         | 1860-70     |                |                  |      |                  |                  | 95 (88.7,98.3)   | 94 (87.4,97.8)     | -                   | -    | -                |
|            |         | 1834-44     | 82 (73.1,89.0) | 88 (80.0,93.6)   | -    | -                | -                | -                | -                  | -                   | -    | -                |
|            | GB      | 1901-11     |                |                  |      |                  | 74.5 (67.9,80.4) | 76.5 (70.0,82.2) | 94.5 (90.4,97.2)   | 90.5 (85.6,94.2)    |      | 87.5 (82.1,91.7) |
|            |         | 1860-70     |                |                  |      |                  |                  | 90.5 (85.6,94.2) | 93 (88.5,96.1)     | -                   | -    | -                |
|            |         | 1834-44     | 74 (67.3,80.0) | 91 (86.2,94.6)   | -    | -                | -                | -                | -                  | -                   | -    | -                |
|            | CFM     | 1901-11     |                | 94 (90.7,96.4)   | -    | 95.3 (92.3,97.4) | 93.3 (89.9,95.9) | 95 (91.9,97.17)  | 93.6 (20.29,96.14) | 92.97 (89.46,95.61) |      |                  |
|            |         | 1860-70     |                |                  | -    |                  |                  | 94.5 (90.4,97.2) | -                  | -                   | -    | -                |
|            |         | 1834-44     |                |                  | -    | -                | -                | -                | -                  | -                   | -    | -                |
|            | CF      | 1860-70     |                |                  |      |                  |                  | 83 (77.1,88.0)   | 96.5 (92.9,98.6)   | -                   | -    | -                |
|            |         | 1834-44     | 55 (47.8,62.0) | 85.1 (79.8,90.1) | -    | -                | -                | -                | -                  | -                   | -    | -                |
|            | CM      | 1860-70     |                |                  |      |                  |                  |                  |                    | -                   | -    | -                |
|            |         | 1834-44     |                |                  |      | -                | -                | -                | -                  | -                   | -    | -                |

\* Links to later target censuses include a number of mistranscribed births, where event year is transcribed as 20+ years before actual birth year.

+ There sample (100 cases) includes the total number of links to that census

## 6.4 Presentation of links in Link-Lives release 2

Release 2 includes the links file created for release 1 (`links_v1.2.csv`)<sup>32</sup> and release 2 (`links_v2.1.csv`), including only machine learning links. For both of them, we display one link per row, with information of the person appearances that are linked, the method used to create them as well as some specific variables relevant for each type of approach and release. See section 12.8 for the codebook of both files, specifically tables 49 for rule-based links and table 50 for the new machine learning links.

Both files include a variable **method\_id**, which can adopt the values in table 32.

---

32. The links and data are the same but the terminology for the links has been adjusted for ease of interpretation to refer to “origin” and “target” sources instead of sources 1 and 2

Table 32: Description of method ids used in Link-Lives version 2.

| method_id | Type             | Name                   | Description  |
|-----------|------------------|------------------------|--|
| 0         | Rule Based       | RB-Primary             | Rule-based algorithm based on invariable information (age, sex, name, birth place) and household support for disambiguation (see section 6.2). |
| 1         | Rule Based       | RB-Household           | Rule-based algorithm based on household matching with a person-appearance linked with the RB-Primary method (see section 6.2).                 |
| 2         | Manual           | DE                     | Computer Assisted Domain Expert Approach (see 6.1).  |
| 3         | Machine learning | C0                     | XGB algorithm developed for censuses without birthplace information (1787-1845)  |
| 4         | Machine learning | census with birthplace | XBG algorithm developed for censuses with birthplace information (1845-1901)   |
| 5         | Machine learning | CFM                    | XBG algorithm developed for linking three person appearances (child, father and mother) from parish registers to censuses (see section 6.3.2). |
| 6         | Machine learning | CM                     | XBG algorithm developed for linking two person appearances (child and mother) from parish registers to censuses (see section 6.3.2).           |
| 7         | Machine learning | CF                     | XBG algorithm developed for linking two person appearances (child and father) from parish registers to censuses (see section 6.3.2).           |
| 8         | Machine learning | GB                     | XBG algorithm developed for linking two person appearances (groom and bride) from parish registers to censuses (see section 6.3.2).            |
| 9         | Machine learning | 1P                     | XBG algorithm developed for linking one person appearance without any household support from any source to censuses (see section 6.3.2).       |

All links from release 1 are included in life-courses but some of them appear in more than one life-course, as identified in **duplicates**. However, not all links created and described in release 2 are used in life-courses, as described in section 7.2.1. That information has been stored in the variable **in\_lifecourse**, a True/False indicator.

Table 33 summarizes the number of links available in the two files.

Table 33: Number of links in release 1 and release 2

| Method                                    | Release 1 | Release 2 |
|---|-----------|-----------|
| Rule Based – RB-Primary                   | 5,427,379 | –         |
| Rule Based – RB-Household                 | 47,700    | –         |
| Manual – DE                               | 24,310    | 25,967    |
| Machine learning – C0                     | –         | 1,531,945 |
| Machine learning – census with birthplace | –         | 2,384,528 |
| Machine learning – CFM                    | –         | 8,375,362 |
| Machine learning – CM                     | –         | 1,052,778 |
| Machine learning – CF                     | –         | 1,101,336 |
| Machine learning – GB                     | –         | 3,203,192 |
| Machine learning – 1P                     | –         | 626,784   |

## 7 Life course aggregation

Life course aggregation consists of a series of steps carried out to extract meaningful life-courses out of all the links between two sources created by our models. A life course consists of anywhere from two records connected through one single link up to many records from different sources, connected and interconnected in complex patterns by multiple links from different models. Over the course of the project, we have developed different methods for aggregation.

### 7.1 Self developed algorithm for rule-based record linkage in release 1 (also included in release 2)

For release 1 we developed our own hard-coded approach for the limited amount of links produced by the rule-based method. This description is also largely the same as that in the guide release 1. We keep both the links as they were published and the text largely as it was first published.<sup>33</sup>

The overall strategy was to identify “end points”, i.e. records which are only connected in one direction, then build the life course by connecting the links iteratively backwards. Some sources, like the 1901 census, are linked to multiple sources (the 1885 and 1880 censuses). In those cases, a life course can “branch”. This can cause multiple, highly similar life courses which can have several links in common. All of these were part of release 1. The method created more than 4 million life courses with up to 12 person appearances.

### 7.2 Life-course aggregation method for machine learning links in release 2

The increased number of links between parish records and censuses created longer and more complex life-courses. Sometimes this produced improved results, for instance, as parish records could be linked to all censuses, they could act confirming relationships between records, i.e., the birth of a person linked two census records from different years, which in turn were linked to each other. Other times, it could create completely impossible life-courses, if two linked census records from different years were in turn linked to two different births. In order to ensure the most reliable data, we developed a more systematic approach to keep the most likely life-courses. Some of the steps in this process were performed at the level of the links and some at the level of life-courses.<sup>34</sup>

#### 7.2.1 Selection of high quality models for aggregation

We decided to select only a subset of all the new machine learning models and ranges described in section 6.3.2 based on the results of the external validation we carried out (see section 6.3.7).

---

33. The rule-based life-courses from release 1 can be found in the file “life\_courses\_v1.2.csv”.

34. The machine learning life-courses from release 2 can be found in the file “life\_courses\_v2.1.csv”.

### 7.2.1.1 General approach to model selection

The overall approach derived from the empirical results described in section 6.3.7 and the focus was to implement a conservative selection to ensure high quality of life-courses. The empirical base for this decision was simultaneously substantial at the collection level (15.000 validated links) but very limited for each model/event/chronology combination. We did not count with thorough tests for all the possible combinations between sources, models, chronologies and geographies, and the size of the samples for the combinations we did we did carry out was very limited (100 persons or events), creating large confidence intervals.

A visual inspection of the tables revealed that some models performed substantially better than others and it was easy to see the differences between the best and worst combinations. However, deciding specific thresholds to decide which model combinations were deemed good enough for life-course aggregation proved challenging. So we develop a set of rules of thumb to systematically take decisions based on the evidence from the validation results, developed by a team discussion, followed by the codification of the decisions by project leaders. A limited number of micro-decisions were also necessary at the moment of implementation. The following rules of thumb were defined:

- We used the confidence intervals in tables 29, 31 and 30 and we took decisions model by model and for the chosen chronologies (before/after 1845 mostly). And when models linked different censuses, we examined how well they performed after/before the event.
- Overall, models were considered to perform “well” (and thus, selected for life-course aggregation) if the lower confidence interval was above 90%.
- However, if there was only one case of one combination model-chronology-event-census performing well while the rest were not, the whole model or model-source combination (the full line in table 31 was discarded.
- Similarly, if a model seemed to perform well over several censuses and the lower threshold of one of the combinations was slightly over the 89%, the suboptimal combination was also kept.
- We decided whether models were accepted for linking to one, two, or more censuses after/before the event, extrapolating directly what we can see from our patchy empirical evidence. Given the different distance between our censuses (5, 10 or 20 years), this means that links are kept for a given model differently over the whole 19th century. However, we prefer this approach instead of the arbitrary decision of deciding on an “x” number of years before or after, as there was no empirical evidence to support any particular number.

### 7.2.1.2 Specific models considered to perform well enough to include in life-course aggregation

- **Census:** model C (links between censuses 1845-1901)

- **PR baptisms:** model CFM linked to the next two censuses after **birth\_year** of the child (three censuses if the birth is on a census year)
- **PR marriages:** model GB linked to the next census year after **event\_year** (two censuses if the marriage is on a census year).
- **PR confirmations:** model CFM with a varying number of censuses depending on **event\_year**:
  - If **event\_year** is higher or equal to 1845, two censuses (three if confirmation is on a census year)
  - If **event\_year** is lower than 1845, one census (two if confirmation is on a census year)
- **PR burials:** model CFM when **event\_year** was higher than 1892
- **Copenhagen burials:** none

### 7.2.1.3 Results

As a result, we created a subset of links to be used for life-courses with only optimal cases. In practical terms, we omitted entirely some models entirely (C0, CM, CF and 1P) and selected links from the original set of links (see table 34). These links are still available within the file “links.v2.1.csv” so researchers can explore whether they can be used in other ways.

Table 34: Summary of omitted links by event type, model, and reason

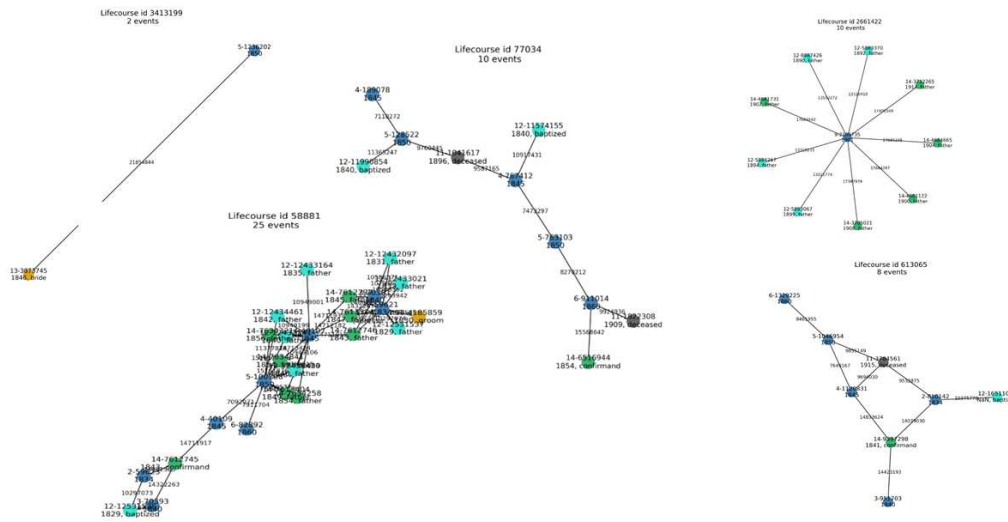
| Event         | Model | Original  | Reason for omission  | Nr. omitted |
|---------------|-------|-----------|--|-------------|
| Baptisms      | CFM   | 3,660,769 | Linked backwards in births                                     | 378         |
|               |       |           | Linked to census after the next two in births                  | 423,558     |
| Marriages     | GB    | 1,836,382 | Missing <b>event_year</b>                                      | 8,582       |
|               |       |           | Linked backwards   | 24,410      |
|               |       |           | Linked to census after the next one                            | 888,010     |
| Confirmations | CFM   | 3,963,587 | Missing <b>event_year</b>                                      | 178,112     |
|               |       |           | Linked to census after the next one (>1845) or next two (1845) | 98,256      |
| Burials       | CFM   | 751,006   | Missing <b>event_year</b>                                      | 342         |
|               |       |           | Event year before 1892   | 41,887      |
|               | GB    | 1,366,810 | Sub-optimal model  | 1,366,810   |

Overall, we kept only 10,669,209 out of the 18,301,892 original links we had created.

### 7.2.2 Initial life-course aggregation

We used the Python library “NetworkX” with the selection of the machine learning links to construct life-courses, that converts easily links into life-courses. This approach allows us to create network plots for each life-course that can help with examination and ease its analysis. Figure 32 shows examples of life-courses. Every life-course is presented as a network, where each circle represents a **pa\_id** from a given source and has been labelled and coloured with some information to ease interpretation. The lines between circles represent links, with the **link\_id** printed on top. These five examples illustrate the varying level of complexity they may show and the different level of interconnection. In the first example, each record is only connected by a line to the next, so it is a simple aggregation. However, in the right bottom life-course, triangles and squares show multiple records are connected to each other, lending further strength to the consistency on the life-course.

Figure 32: Examples of visualization of life-courses



### 7.2.3 Life-course level sanity checks

After creating an initial set of life-courses using the selected models and links, we implemented a sequential set of sanity checks to exclude life-courses with conflicting information. This is a very conservative approach, given that in many cases, only one of the links introduced the inconsistent information. It would be technically possible to try to identify the person appearance with the inconsistent information and remove it. However, given the resource constraints of the project, it was not feasible to implement and test any complex identification of inconsistent person appearances. We maintained instead our conservative approach in keeping fewer coherent life-courses. As the unused links are still available within the links file, researchers can still access them and use them themselves to improve our



release 2 life-courses. That is possible for recovering additional life-courses but also adding additional person appearances to the life-courses we have created. For instance, given that we needed to omit most of links to burials and deaths, life-courses have rarely a person appearance with a death associated. However, it is possible to merge the life-courses and the link files and determine rules for inclusion of person appearances of deaths that are consistent with the existing life-courses.

We removed entirely life-courses if a life-course contains:

- More than one person appearance in a single census year (ie, two person appearance of the 1860 census).
- More than one person appearance from baptisms as a main person<sup>35</sup> on different dates<sup>36</sup>. Two baptisms are allowed in a life-course if they happened on the same date to account for the existence of duplicate parish records (i.e, for instance, double recording in regular residence and transient place).
- More than one person appearance from confirmations as a main person.
- More than one person appearance from burials in Parish records as main person on different dates<sup>37</sup>.
- More than one person appearance from burials in Copenhagen Burials as main person on different dates<sup>38</sup>.
- A person appearance preceding a person appearance from baptisms as a main person (if a main person baptism person appearance is present, it has to be the first event)
- A person appearance following a person appearance from burials in either the parish records or Copenhagen burials as a main person (burials as main person can only be last events. However, if the **role** in the burial is “mother”, it can be after the main person’s death.
- A person appearance from marriages before a person appearance from confirmation for a main person.
- A person appearance from marriages before age 14 compared to birth,( i.e., less than 14 years between birth and marriage event, allow 14 exactly but discard anything below).

---

35. Main persons are defined according to **role**: For censuses: All records are main (role is not filled out). CBP: deceased is main (no other roles in that source). PR burial: deceased is main (spouse, father, mother are secondary).PR birth: baptized is main (mother, father are secondary).PR marriage: bride and groom are main (groom-mother, groom-father, bride-father are secondary). PR confirmation: confirmand is main (father and mother are secondary). See more on this in section 6.3.2

36. “Different” is encoded as any deviation on day, month or year

37. Same definition as above

38. This is not actually implemented as no Copenhagen burial person appearances have been included in lifecourses

- A person appearance from confirmations before age 13 compared to birth ( i.e., less than 13 years between birth and confirmation event (allow 13 exactly, but discard anything below).
- A person appearance from confirmations after age 25 compared to birth, (i.e., more than 25 years between birth and confirmation event but allow for 25 exactly).
- More than one gender if at least one person appearance is missing gender. This check is intended to catch lifec-ourses where two persons have been mixed on the account of the sex being missing in an event. Hence, we allow a life course with multiple sexes since the model deliberately chose to make these links (but there should still be none of those since we blocked on sex).

The implementation of the life-course sanity checks meant the removal of a total of 23,947 life-courses, less than 1%. The largest group was that of life-courses where the aggregation connected two different person appearances from the same census. Figure 33 shows an example of this case. We can see that the person appearance of a woman at the confirmation of a child in 1861 (marked in green, “14-19842208”) is connected to person appearances in two different censuses: 1850 and 1860 (in dark blue and circled in red). However, additionally, the 1850 census person appearance has been further linked to a different person appearance from the 1860 census.

Another case is shown in figure 35, where the sanity check is not passed because there are two person appearances from births as a main person (circled in red). An analysis of the life-course reveals that the death (highlighted with a yellow square) connects two unrelated (and consistent smaller life-courses). In these two examples, as in many others that do not pass the sanity checks, it is somewhat straightforward for a domain expert to decide which person-appearance to remove or how to divide a large “insane” life-course in two smaller “sane” life-courses. However, we did not have resources to implement any systematic identification and processing of these cases so we discarded the full life-courses and are not presented in the life-course file.

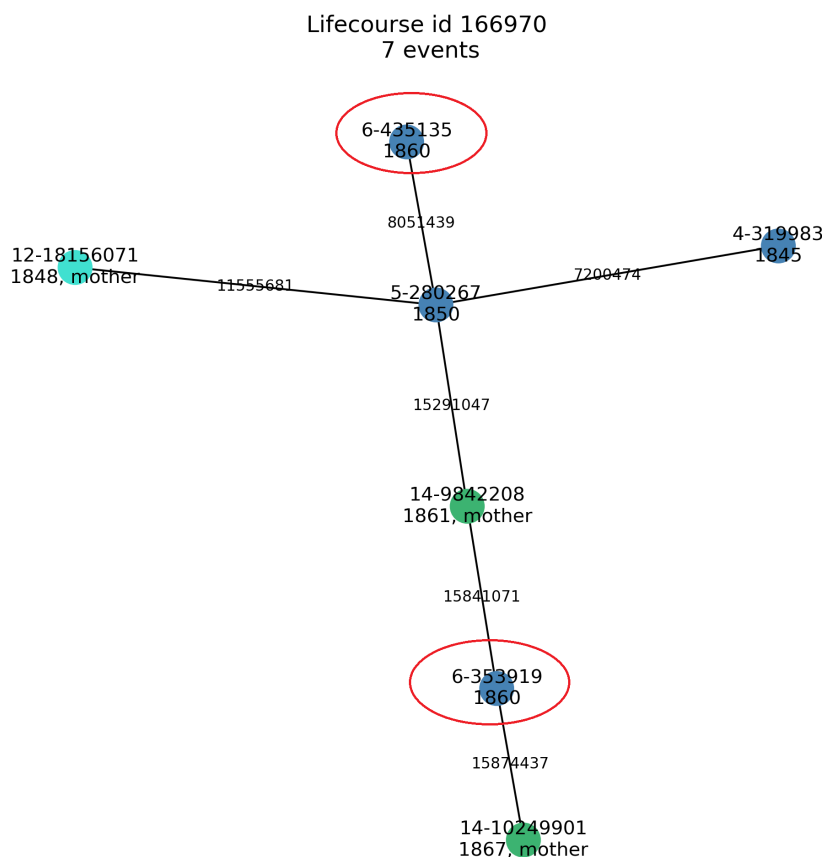
#### 7.2.4 Results

The overall process has yielded 3,548, 215 life-courses after the implementation of the sanity checks. Figure 35 shows the distribution. A substantial amount (60%) consists on two person appearances connected by one link but there are also fair amount of life-courses with more person appearances. The maximum amount of person appearances linked in a single life-course is 36.

#### 7.2.5 Life-course quality test

We have performed a minimal quality test to assess the internal coherence and plausibility of the life-courses after the implementation of the model selection and sanity checks. We drew a sample of 50 records and, given that there was a large proportion of smaller life-courses and we wanted test all types, we stratified them by size in three groups (small, 2-4 linked records; medium, 5-7, and large,

Figure 33: A life-course with two person appearances from the 1860 census



more than 8). For the task, we only used single domain expert for validation (three team members were involved) who used the life-course information in our dataset, life-course plots as those shown in figures 34 and 33 and the inspection of person appearances in the Link-Lives webpage. Out of the 50 life-courses validated, we only found 4 with one wrongly linked person appearance, which provides a precision rate of 92%. The sample is so small that we cannot draw statistical conclusions but the results did not show an over-representation of large life-courses among the cases with problematic links. Additionally, we found that the four cases could be easily prevented with further work in our models, given that the problems arise from mismatch in places of birth, whose standardization and overall treatment are sub-optimal. We have not performed detailed analyses of the difference between releases 1 and 2 models but we consider the newest models vastly superior. Machine learning models have higher coverage and we have designed more strict selection rules to ensure high precision. Additionally, our validation of life-courses has shown that we find many life-courses also found

Figure 34: A life-course with two person appearances in births as main person

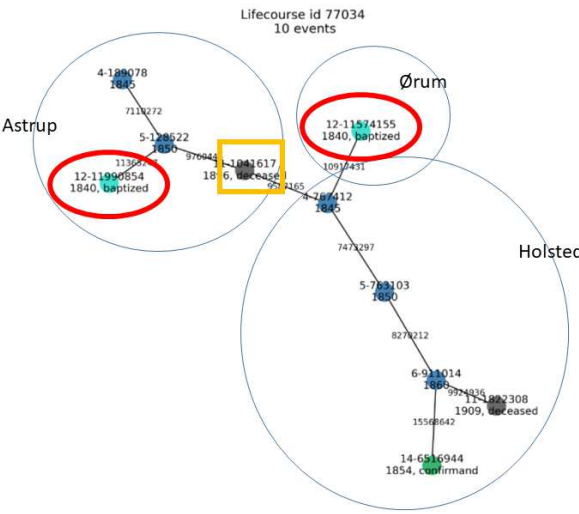
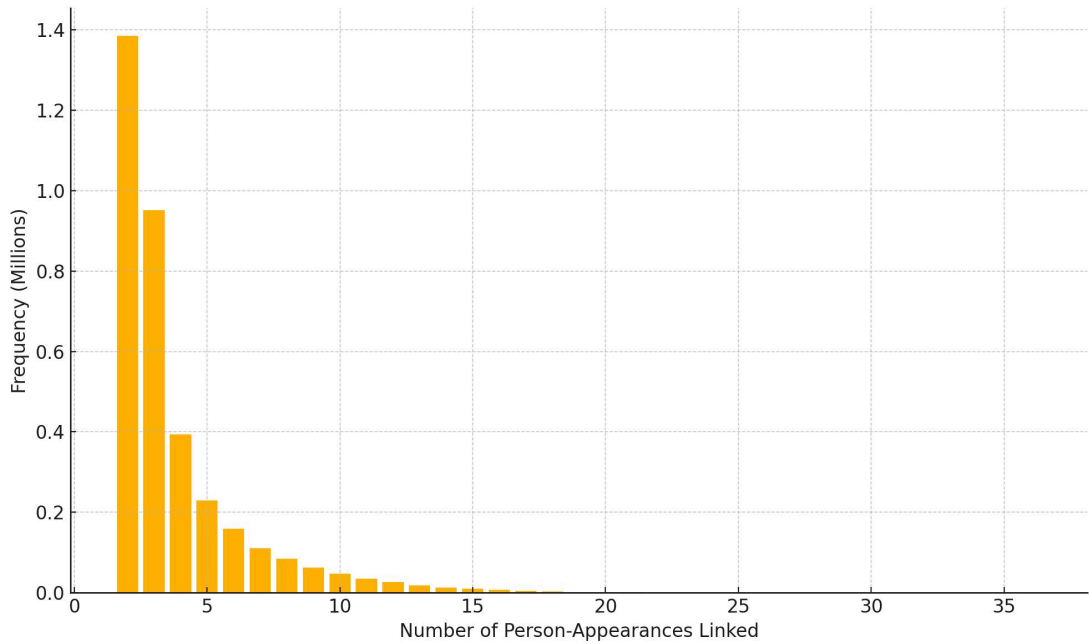
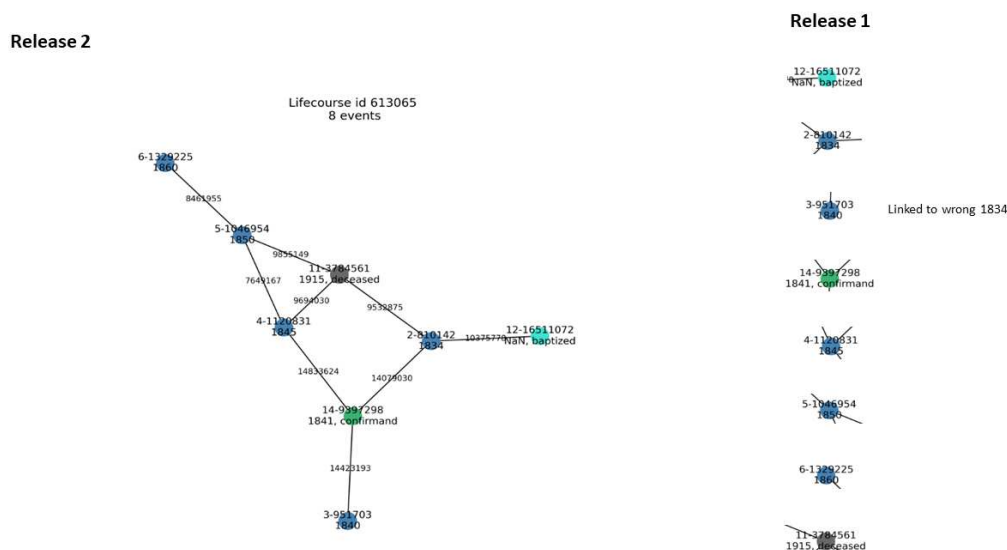


Figure 35: Size of life-courses



by the rule-based models (specially on censuses) but we have gained a substantial amount of more complex life-courses, including parish records. Figure 36 shows an example of a whole new life-course that can be found by the new approach that was not at all available at all in release 1.

Figure 36: Example of improvements in release 2



### 7.3 Presentation of the life-courses in Link-Lives release 2

In release 2 we provide two life-course files, pertaining to the life-courses released in release 1 and the new life-courses from release 2: “life\_courses\_v1.2.csv” from release 1 contains only life-courses from the rule-based algorithm and “life\_courses\_v2.1.csv” contain the life-courses created for release 2. The file from release 1 contains 3,138,251 life-courses and the file from release 2 contains 3,548,215 life-courses.

Both files have the same the structure (see section 12.9 and see table 51 for more details) but there is a difference in how to use the variable **link\_ids**. All person appearances associated with one life-course are stored in the most efficient way, as one line per life-course.<sup>39</sup> Each record is a life course, with a unique **life\_course\_id** and there are three variables that contain all the information about what links it is comprised of (**source\_id**, **pa\_id** and **link\_id**), separated by commas (“,”) and a final number of variables that indicates the number of sources involved.

Figure 35 shows an example of the content for release 1. In **life\_course\_id** number 4, the information on the related person appearances is divided between **\_sources** and **pa\_ids** given that unique identifiers can only be created by combining both. There are three **pa\_ids**, whose respective **source\_ids** are listed in the respective columns in sequence. That means that this life-course is comprised of **pa\_id** 1000002 from **source\_id** 3 (census of 1845), is linked by **link\_id** 2360648 to **pa\_id** 707136 of **source\_id** 4 (census of 1850). The links that connects these person appearances are listed in order in **links\_ids** and **\_n\_sources** provides a summary of the number of person appearances involved. The only difference for

39. This structure is efficient for storing and dissemination but not ideal for analysis so we provide some suggestions on how to work with them in section 8.

data from release 2 is that the variable **links\_ids** contains only a subset of the **links\_ids**, only as far as they fit the sequence from the sources. As release 2 has links in multiple directions, it is impossible to capture them in the current system, so for information on these, we recommend retrieved the **links\_ids** associated to the **pa** in the `links_v2.2.csv` file

Table 35: Example of a life course in life-courses file.

| <b>lifecourse</b><br><b>_id</b> | <b>pa_ids</b>                   | <b>source</b><br><b>_ids</b> | <b>link_ids</b>     | <b>_n</b><br><b>sources</b> |
|---------------------------------|---------------------------------|------------------------------|---------------------|-----------------------------|
| 4                               | 1000002,<br>707136,<br>731392   | 3, 4,<br>5                   | 2360648,<br>1882979 | 3                           |
| 7                               | 5299640,<br>1000002,<br>4504898 | 12, 9,<br>14                 | 4561448,<br>5201654 | 3                           |

## 8 Extracting data

This section is a short guide on how to get started with LL data. It covers how the source files are connected to the links and life-course data.

Most data files will be too large to open with Excel or any similar software for spreadsheets. We recommend to use a program that can deal with millions of records, e.g. Stata, R, Python or any other programming language suitable for data analysis. We recommend that you work on a computer with at least 16 GB RAM. It is possible to work with lower memory but risk anything from a slow response to crashing the computer.

### 8.1 How to combine harmonised and original transcribed datasets

In order to combine the harmonised data (files ending in “.cl”) and transcribed datasets (files ending in “.std”) you need to load the two versions of a given source, for instance, the original transcription of 1880 census (“census\_1880\_v1.cl.csv”) and the LL harmonised file (“census\_1880\_v1\_std.csv”) and use the variable **pa\_id** for merging them. As each **pa\_id** is the unique ID variable within a source (in this case, the 1880 census) you do not need to specify the source. This process holds for any of our sources (censuses, parish registers or Copenhagen burial registers).

### 8.2 How to combine links and LL harmonised datasets

In order to combine person appearances from two sources with the help of the links file, you need to use two variables: the **pa\_id** that identifies each record (person appearance) within a given source and an id for the source, called **source\_id**. The combination of both makes a record unique across all Link-Lives datasets. All source files (both original transcriptions and harmonised files) have a variable with **source\_id** to ease merging.

When combining two sources, for instance the 1860 census with the preceding census of 1850, you need to perform two merges in succession:

- Merge the harmonized 1860 census to the file containing your desired links (“links\_v1.2.csv” or “links\_v2.1.csv”) using the combination of **source\_id** and **pa\_id** as key. You have to merge the census **source\_id** and **pa\_id** with the origin variables in the links files, those suffixed with “.o”: **source\_id\_o** and **pa\_id\_o**.
- Merge the result with the harmonized census of 1850. You have to merge the census variables **source\_id** and **pa\_id** with the target variables in the links files, those suffixed with “.t”: **source\_id\_t** and **pa\_id\_t**.

We recommend that you perform a left merge from origin to the links files (for instance, from 1860 to links) to ensure that your population matches that of your starting dataset. You should decide if you would rather perform a right merge from that resulting dataset to the 1850 if you only want the linked individuals or

an outer merge if you would like also to keep the unmatched individuals in 1850. That will depend on your research question.

### 8.3 How to combine life-courses and LL harmonised datasets

The release format of the life-courses is not suited to direct analysis so we suggest to transpose the information from one life-course per record to one person appearance/event per record. This structure will be better suited for analysis, once the data has been combined with information from the source datasets.

We can see an example of how life-course no. 4 from the life-course file in table 35 would look transposed in table 36, where each of the person appearances have been made into a record.

With this new format, it is relatively straightforward to merge successively the different harmonized datasets to the values of **source\_id** and **pa\_ids** to reconstruct the datasets.

The person appearances connected to any given life-course are not ordered in any chronological way (only by **source\_id** which is not meant to be chronological). The resulting set here is chronologically order because sources 3, 4 and 5 correspond to the censuses of 1845, 1850 and 1860. However, when including parish registers and other sources, this order will be broken as e.g. parish register baptisms has a **source\_id** of 12 which will most likely always place it in the last position.

Thus, we would recommend to sort through the variable **event\_year** after merging this file sequentially to different harmonized datasets and creating a variable with the lifecourse order.

Given the size of the datasets, it may not be possible to merge all harmonized datasets to this transposed file, so it may be more efficient to select some sources to do so or do it sequentially.

Table 36: Transposed life-course no. 4.

| <b>lifecourse_id</b> | <b>pa_ids</b> | <b>source_ids</b> | <b>n_sources</b> |
|----------------------|---------------|-------------------|------------------|
| 4                    | 1000002       | 3                 | 3                |
| 4                    | 707136        | 4                 | 3                |
| 4                    | 731392        | 5                 | 3                |



## 9 Documentation on the construction of auxiliary datasets

### 9.1 ALA versions of data

We created a specific version of datasets to be used within our software ALA for computer-assisted manual linking (see section 6.1.3). These versions merged variables from the transcribed datasets (files ending `*_cl.csv`), from the harmonized (`*_std.csv`) and included a minimal amount of variables created specifically for manual linking. The corresponding datasets for censuses and Copenhagen burials are part of the release under the folder `main_datasets/ALA`, while the equivalents for parish records can be requested along with the rest of parish records (see section 2.7).

#### 9.1.1 Censuses

There is one ALA file for each census and the ALA versions of the 1787-1901 censuses consist on 17 variables. The codebook is included in section 12.5, table 46. It includes the mapping between the variables available in ALA and the variables available in the transcribed and harmonized datasets.

#### 9.1.2 Parish registers

The ALA versions of parish registers do not mirror the structure of transcribed and harmonized versions. As the files for each of the events are really large, it was not feasible to load them in ALA so they were divided into parishes. They include 20 variables, not all of them actively used in the linking display we used. In section 12.6 we present the codebook, table 47. It includes the mapping between the variables available in ALA and the variables available in the transcribed and harmonized datasets.

#### 9.1.3 Copenhagen Burial Registers

There are four ALA files of Copenhagen Burial Registers, breaking down the period 1861-1911 to reduce the need to include the full datasets. They include 24 variables, not all of them actively used in the linking display we used. In section 12.7 we present the codebook, table 48. It includes the mapping between the variables available in ALA and the variables available in the transcribed and harmonized datasets.

### 9.2 Names

We standardized names using a name synonym catalogue that we created for the purpose. It was created to ease the work of linking individual information from historical sources and reconstruct life-courses. The problem it tackles is the lack of spelling consistency in person names across different sources. Vick and Huynh (2011) have shown with American and Norwegian examples that standardization helps linking algorithms.

Our approach has used a conservative approach to reducing spelling variations that can refer to the same name. The file we make available on this release, `SC_names.v1.csv`, contains standardized versions of a selection of single names extracted from the full onomastic profiles (that is, the full name with all its potential onomastic components, names, patronyms, family names) of the Danish national censuses of 1787, 1801, 1834, 1840, 1845, 1850, 1860, 1880 and 1901. See more details on the structure of file we make available in section ???. The data used to create it stems from the same extractions used for the rest of Link-Lives (see section 2.3). Existing synonym catalogues, created by team members that had worked on similar topics, were used as a starting point, from where we developed a new approach.

### 9.2.1 Synonym catalogue by Thomsen

An early synonym catalogue that served the same purposes was the standardization tool created in 2007 on the basis of an extract of 11,325,950 full names from the transcribed Danish censuses in *Dansk Demografisk Database* by Link-Lives team member Asbjørn Romvig Thomsen in the context of his PhD (Thomsen 2010)). Only censuses from the Danish speaking areas (i.e. the Kingdom of Denmark borders 1864-1920) were included. These full names were split into a maximum of four parts: the first word, the second word, the third word and then the rest - if any - of the name string (even if consisting of more than one word). After discarding the strings with more than one word and non-words (numbers or signs), the total of unique name words amounted to 306,243. Every unique name appeared on average 88 times - the vast majority much less and a minority much more often: The most common name word was “marie” which appeared 874,405 times.

The 5,342 most frequently appearing single names - and some 500 less frequent ones which had been standardised as part of a pilot project were standardised manually. It resulted in a coverage of 95% of the total number of name appearances.

The manual standardisation process was based on the creator’s expert historical knowledge of name spelling practices in official documents ca. 1750-1850, especially in Danish rural areas. The standardisation was not based on philological views, but rather on practical experiences with locating individuals across different historical sources. The purpose was pragmatic: the standardisation was meant to be a practical tool (a) for searching for individuals in the census transcriptions and (b) for linking individuals between censuses.

### 9.2.2 Synonym catalogue by Revuelta-Eugercios and Kællerød

The synonym catalogue used the full onomastic profiles for the census of 1880 only. Their work was aimed at identifying onomastic components used as middle names (the occurrence of names immediately before the surname) even before the use of that concept. It was used for an article (Kællerød and Revuelta-Eugercios 2015), which describes the methodology in full and its results used in a PhD Thesis (Kællerød 2019). We describe briefly here the features relevant for its contribution to Link-Lives developments. After cleaning non-desired elements (non alfa-numerical characters, expressions that did not contain names as titles

or descriptions “baby boy/girl”, etc), they tokenized the full onomastic profiles into single components, names, which were standardised and classified.

Each name was assigned different types for male or female, as for instance the name “Martin” can be both a first name and surname for men but it should only be a surname for women. Thus, the following categories were considered:

- male first name
- female first name
- unisex first name
- initial
- abbreviation
- patronymic
- family name
- female first name and family name
- male first name and family name
- unisex first name and family name
- unknown

To ease the classification work, they extracted list of names from different databases where the division of first names, middle names and surnames from historical sources had been performed manually by humans as well as more modern data sources. They integrated data from:

- The death certificates of the city of Copenhagen for the years 1880–1882 (which were transcribed for another project by one of the authors)
- *Det Danske Udvandrerarkiv*, DDU (‘Migration Records in Denmark’), 1880–1918.
- The synonym catalogue created by Asbjørn Romvig Thomsen (see section [9.2.1](#)).
- Harvested data from *Danskernes Navne* (‘The Names of the Danes’), which contains name information of the approximately 6.5 million Danes registered in the Civil Registration System, CPR, since this was established in 1967 and up to 2005.

These categorizations were augmented and revised manually by Kællerød to produce a near comprehensive typology table, with particular attention paid to any name component appearing 25 times or more within the census of 1880 (equating to 5,700 classifications, accounting for 93.4% of the 5,300,939 naming instances). Moreover, Kællerød standardized further all the names in this list.

### 9.2.3 Methods for creating the Link-Lives name synonym catalogue

At the start of the the Link-Lives project in 2019, the need for an updated tool to standardise names led to the creation of the Names Synonymy Catalogue. The actual standardisation was performed by Asbjørn Romvig Thomsen, Samantha Norholdt Aagaard and Birgit Eggert with the participation of Nicolai Rask Mathiesen and Barbara Revuelta-Eugercios in the design.

The basic datasets for the creation of Version 1 were

- an updated extract of all unique name strings from the censuses
- the two above mentioned synonym catalogues by Thomsen (section [9.2.1](#) and by Revuelta-Eugercios & Kællerød (section [9.2.2](#).

We extracted the full name strings (the full onomastic profile) of all available censuses (1787-1901) and split them into single “names” (understanding “name” as any element included in the full onomastic profile). We used the space between names as a marker of a new name, “ ”. For instance, “ Hans Kristian Jensen” is split into “Hans”, “Kristian”, and “Jensen”.

We decided to standardise 95 % of the name string appearances of the census material, which involve standardizing 6,233 names. Most of them were already standardised in the previous synonym catalogues, but the old standardisations were examined and new ones added. The standardisation went through four stages:

1. Thomsen and Aagaard standardised all the strings independently.
2. They discussed and decided upon cases with different decisions
3. Eggert, researcher in Onomastics, suggested changes to the synonym catalogue from a philological perspective
4. Thomsen and Aagaard decided which suggestions to include from a linking perspective

All single name appearances were linked to a standardised spelling of that name which was considered a common identifier of the spelling variations. For practical reasons, these standardised versions (column: `LL_standard_name`) were spelled with the use of as few and as “simple” letters (like “k” instead of “c” or “q”) as possible to make it easier for the standardiser to remember how to spell the standard version of the name.

It was also decided to note whether the string was:

- not a name (stored in a separate variable, **IkkeNavn**) - not standardised.
- an abbreviation of a name (variable: **Forkortelse**) - not standardised.
- a spelling variation which indicated that the name could be in the genitive case (column: **Ejefald**) - relatively few cases, inconsistently standardised into the standard version of the name + an “s”.

A further classification of names into first names, patronyms, and family names, on the other hand, was not undertaken. Many Danish names are easy to classify into these groups, but there is also a substantial group which can actually be both a first name and a family name (like e.g. “Ernst”). Furthermore, the distinction between the patronym (as derived from the father’s first name) and the family name (as unchanged from generation to generation) becomes anachronistic during the 19th century, as formerly patronymic surnames were increasingly used as unchangeable family names. It was therefore decided that this classification should instead be carried out automatically from the positions of the single name strings in the full onomastic profiles in the sources. At the same time, this more transparent approach would provide the classification for the full dataset, instead of only for the most frequent strings.

#### 9.2.4 Description of the synonym catalogue included as part of the release

The current file we provide, `SC_names_v1.csv`, is a simplification of the fuller synonym catalogue we developed and described above. It only contains two variables, **original** and **standard**, so it can be used for other datasets with names (see also a description of the codebook in section ??). The additional variables described above, as well as additional information and new names we have developed in the last years, can be requested following the same procedure as for any not readily downloadable data request, see section 2.3. We hope to make it available as a downloadable file in the future.

### 9.3 Geography

#### 9.3.1 Fitting the census geography to the existing administrative boundaries: the Danish census historical GIS

The project built a historical GIS to combine the idiosyncratic characteristics of the Danish census transcription with the information on geographic boundaries for Danish administrative units. It has two components: census snapshots for the years 1787, 1801, 1860, 1880 and 1901 that approximate as accurately as possible the census transcription geography by adapting the administrative units in DigDag to represent and analyse a chronologically accurate map. Second, a persistent geography of aggregated invariable units (where smaller parishes have been clustered to create units with unmovable boundaries) for all census years, including 1837 and in order to allow for comparison over longer periods of time. We hope to make it available in the future.

## 10 Benchmark dataset 1845-1901

The Benchmark dataset 1845-1901 is the result of the method described in 6.1, the application of the computer-assisted domain-expert record linkage approach. Its main purpose was to serve as training data/test data for the machine learning algorithms but with a design that also allowed limited research. In section 6.1, we described the main features and link rates. In this section, we describe further elements of its creation.

### 10.1 Structure of the file

The dataset is provided with all the data and metadata created during the process of linking and stored as part in a file called `benchmark_v1.xlsx`. See full information about the content in 10.2. The naming of the variables is slightly different from the main Link-Lives approach for the full datasets and the `links.csv` dataset, see codebook in table 12.11. The reason for this inconsistency is that the process of linking and constructing a finished linked dataset predates the formalization of the Link-Lives standard structure, as described in section 5 and with the codebook in table 12.4. In particular, its origin was tied to how the data was standardized to fit our software ALA (see 10.3).

### 10.2 Content of the file

The file called `benchmark_v1.xlsx` includes both the benchmark dataset for the period 1845-1901 and the benchmark dataset for 1787-1845, whose specifics are described in section 11. It includes 60,957 attempted links for 60,957 person appearances. The two sets can be distinguished using the variable **period**, with the values “1787-1845” and “1845-1901”. There are 2,538 person appearances in the first period and 58,419 in the second period. The dataset includes the outcome of the attempted linking of every person appearance in the chosen linking units, listing the linkers involved (anonymized to IDs), their decisions, as well as metadata on time and versions of the software using the method. That means that the dataset includes also the person appearances that we were not able to link, including the same metadata on them (in contrast with the `links.csv` file where we only show actual links). Additionally, we have added four units linked with a variation of the method, which are identified in the variable **method\_type** = “experiment”, in contrast with the rest of the dataset, which have the value “production”. These units are the result of experiments in which we involved a substantial number of team members for teaching and development of the method with censuses. They involve 3,159 person appearances and our early work. They all have been linked by at least two linkers but in some cases up to five and disagreement solved by a team member. We consider them of as a minimum similar quality of the production files and that is why we add them to the file, so they can be used by others and complete our set of “core parishes” (see section 10.4)

## 10.3 Data used for linking

The data used for linking in our linking interface was a set of specific “ALA versions” of all our transcribed datasets that combined original variables with a small amount of standardised variables to facilitate linking. Depending on the type, these ALA files were broken into smaller segments to speed up loading into the software (i.e, parish records are split in files for each parish) and did not have all available information contained in the transcribed datasets (see full description in section 9.1).

## 10.4 Specific linking units included

In this section, we further describe the details of the specific decisions described in section 6.1.2.1. For censuses, linking units were parishes in rural areas and streets or neighbourhood samples in urban areas, to ensure wide representations of different types of individuals. For parish registers we just used full parishes either urban or rural. Copenhagen Burial Registers was organized chronologically not geographically, so we made samples of roughly 200 individuals in different years with varying numbers of years elapsed from the census.

We initially selected parishes from a sample that historian H.C.Johansen made of representative rural parishes in his book *Befolkningudvikling og familiestruktur i det 18. århundrede* (Johansen 1975, 30–31). We chose some from the 26 listed, focusing on mid-size parishes around 300-500 inhabitants. Afterwards, we expanded the list to parishes outside of his list to cover areas that were missing. We also included samples from urban settings, that were not at all considered in his book.

These units represent islands, rural areas and urban districts including Copenhagen, which was often administered in slightly different ways to the rest of the country. When we chose streets and parishes as samples from larger conurbations, we tried to account for their mixed socioeconomic profiles and different population sizes. Overall we linked data from 200 distinct linking units, which correspond to 75 different geographies (which were parishes, streets/parishes in towns or sample years in Copenhagen burials).

Table 37 shows the actual number of linking decisions, according to the origin source, as our target source is always the census. Thus, “census” means census to census linking, Copenhagen Burials indicates linking that source to the censuses, etc. The highest number of person appearances can be found in censuses, especially the first ones, as they were key to develop our understanding of the process. The second largest number in the table is baptisms but, in fact, as we count person appearances and we typically can have three persons associated to a baptism, they refer to a smaller number of set of events. In figure 37 we show the spread of our dataset. Given the resources available in the project in the early part of the project, we were able to link more distinct areas. However, for parish records to census linking, we secured areas in the Jutland peninsula, the island of Funen and Zealand, where Copenhagen lies.

While we tried to select a variety of contexts, we also tried to create a complete set of linked parishes, we called “core parishes”. We selected seven to link in as many sources as it was feasible but only managed to complete fully three of them.

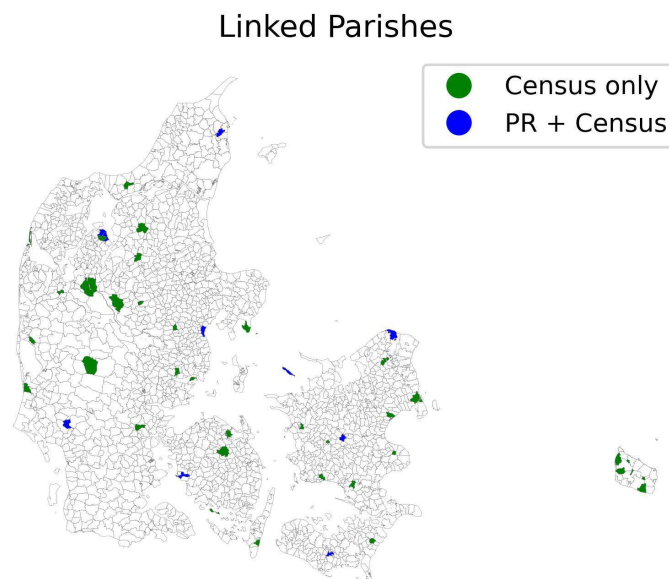


Table 37: Number of person appearances with a decision by origin and target source.

| Origin source | Target census |        |       |        |      |       | All    |
|---------------|---------------|--------|-------|--------|------|-------|--------|
|               | 1845          | 1850   | 1860  | 1880   | 1885 | 1901  |        |
| Census        | 12,010        | 11,037 | 2,993 | 3,494  | 0    | 0     | 29,534 |
| Cph Burials   | 0             | 0      | 1,135 | 1,019  | 688  | 0     | 2,842  |
| PR Baptisms   | 357           | 2,037  | 3,076 | 5,197  | 0    | 4,676 | 15,343 |
| PR Burials    | 881           | 1,950  | 1,231 | 1,437  | 0    | 0     | 5,499  |
| PR Marriages  | 250           | 561    | 923   | 1,559  | 0    | 1,631 | 4,924  |
| All           | 13,498        | 15,585 | 9,358 | 12,706 | 688  | 6,307 | 58,142 |

Our “core parishes” were Aarre, Krønge, Junget, Bringstrup, Kærum, Gærum, Gilleleje and we achieved full coverage on Aarre, Junget and Gærum. See table 38 for a full account. During the linking process, we made several tests with some linking units

Figure 37: Geographical distribution of linking units by origin source



## 10.5 Effects of linking interface on data creation

### 10.5.1 Overall changes in the software and documentation

The linking interface changed as we developed the software to fit the linking needs, new functionalities and new datasets with different structures, information



Table 38: Coverage of origin sources by parish and target census year in the set of “core parishes”

| Origin source | Target census | Aarre | Bringstrup | Gærum | Gilleleje | Junget | Kærum | Krønge |
|---------------|---------------|-------|------------|-------|-----------|--------|-------|--------|
| Censuses      |               |       |            |       |           |        |       |        |
| Census 1850   | 1845          | x     | x          | x     |           | x      |       | x      |
| Census 1860   | 1850          | x     | x          | x     | x         | x      | x     | x      |
| Census 1880   | 1860          | x     | x          | x     | x         | x      | x     | x      |
| Census 1901   | 1880          | x     |            | x     |           | x      | x     | x      |
| PR Baptisms   |               |       |            |       |           |        |       |        |
| PR Baptisms   | 1845          |       |            |       |           | x      |       |        |
| PR Baptisms   | 1850          | x     | x          | x     | x         | x      | x     | x      |
| PR Baptisms   | 1860          | x     | x          | x     | x         | x      |       | x      |
| PR Baptisms   | 1880          | x     | x          | x     |           | x      |       | x      |
| PR Baptisms   | 1901          |       | x          | x     |           | x      |       | x      |
| PR Marriages  |               |       |            |       |           |        |       |        |
| PR Marriages  | 1845          |       |            |       |           |        |       |        |
| PR Marriages  | 1850          | x     | x          | x     | x         | x      | x     | x      |
| PR Marriages  | 1860          | x     | x          | x     | x         | x      | x     | x      |
| PR Marriages  | 1880          | x     | x          | x     | x         | x      | x     | x      |
| PR Marriages  | 1901          | x     | x          |       | x         | x      | x     | x      |
| PR Burials    |               |       |            |       |           |        |       |        |
| PR Burials    | 1845          | x     |            | x     |           | x      | x     |        |
| PR Burials    | 1850          | x     |            | x     |           | x      | x     |        |
| PR Burials    | 1860          | x     |            | x     |           | x      | x     |        |
| PR Burials    | 1880          | x     |            | x     |           | x      | x     |        |

and quality, There were eight releases of the software ALA that were put into production during the period between March 2020 and July 2023. New releases were only introduced in order to incorporate entirely new datasets pairings but they often also improved general aspects that had to do with improvements in functionality. Each new ALA release was created in cooperation between data scientists and domain experts based on thorough examination of the sources, the possibilities of linking they provided (which variables were available, how could they be used in linking, etc.) and the concrete assessments of the needs of the project. Findings from this examination were recorded in Scouting reports, documents that collected characteristics of the sources, the transcription and consideration related to linking and were used for discussion.

Each software release was accompanied by a “User Guide” for the software and a “Best Practices” document. User guides explained to the rest of the team new functionalities, keeping only to the technical details, and Best Practices dealt with how to apply the software to each particular dataset (see more in section 10.6).

### 10.5.2 Stability of the visualisation of variables

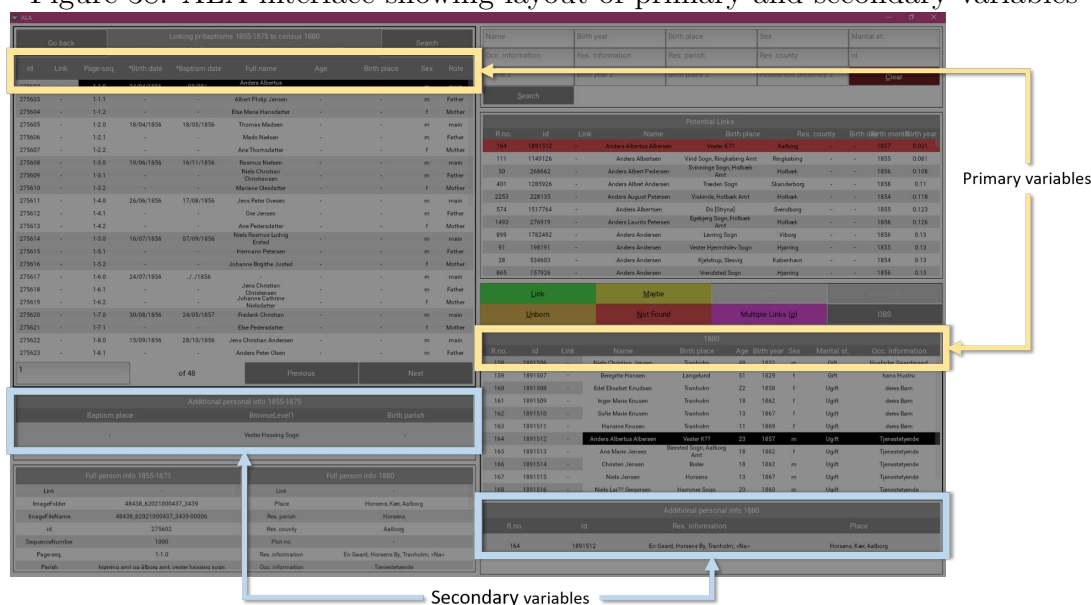
The main description of the software and functionalities can be found in 6.1.3. In this section, we focus on the specifics of display over time. The main display elements have remained mostly unchanged or only changed in so far it was necessary to accommodate slightly different data. The main aspect of the visual layout that has been stable is how we showed variables for individuals in the two

sources. As we evolve to distinguish the use of primary and secondary variables in linking, we also visually showed them differently.

The primary variables used were: full name, age or date/year of birth, birth place, gender, civil status and their co-residents or other individuals that appear in the event they participated in. All other variables were considered secondary, but most commonly we used place of residence and occupation. All variables were listed in the interface, but these were not immediately visible.

We displayed primary information visibly in the main panel (see 38) but secondary information was tucked away separately below to encourage linkers to link first on the primary variables and only secondly use the secondary variables. Users had to scroll down to access that information but also a copy of all information from the original source.

Figure 38: ALA interface showing layout of primary and secondary variables



### 10.5.3 Availability of potential candidates

A feature that may have affected linking is the presence of potential candidates to choose from generated by the algorithm.

The possibility for potential candidates was present from the start to ALA v1.10 and required the production of the candidates through computations outside of ALA that were exported to a file that was made available as part of the data foundation for ALA. Up to 10 potential links per person appearance were calculated on the following criteria and presented in descending order of proximity from a 0 score, the highest possible match, to 0.5, the lowest possible score:

- Blocking: **Age** (+/-2) and **Gender**. In cases where **Bbirth year** was available, this was used instead of **Age**
- Score: combined distance score of **Birth Parish** + **County**, **First Name**, **Patronym** and **Family Name**.

Due to the workload for the ALA v.1.38 and v.1.6, no potentials were produced so linkers had to rely on the search boxes. Starting in ALA v.1.8, the potential candidates were generated dynamically through ALA programming mimicking the parameters above.

## 10.6 Summary of Best Practices

We wrote Best Practices documents based on prior working documents created in the context of developing a new version of ALA and understanding how to link but also heavily depending on the previous version, highlighting changes at the start of each document (from Best Practices nr. 4). Table 39 describes the highlights from each new version. We list the Best Practice number, date and number of the ALA release associated to it, the new linking decisions available for linkers, new functionalities and recommendations changes. We also provide their full text in the accompanying **Link-Lives Paradata** documentation.

Overall, we also stated our four guiding principles at the beginning of the project, which have stayed the same since the first best practices:

- Only link one person to one person.
- Always challenge potential links.
- Always search for competing candidates.
- When in doubt, do not link.

Additionally, we expanded to two further items in Best Practices nr.4:

- Prioritize primary variables before considering secondary variables.
- Do not use unauthorized sources outside of ALA.

Over the course of the project, there were specific changes introduced in the best practices that brought new considerations or a re-thinking of some of the actions, partially brought about by the new datasets, new challenges and the new functionalities.

Table 39: Summary of the Best Practices main changes.

| Nr. | Date & ALA version                  | Sources covered                                  | New link decisions             | New functions  | Recommendation changes  |
|-----|-------------------------------------|--|--------------------------------|--|---|
| 1   | 05-03-2020; ALA 0.3.2, 0.4.1, 0.4.6 | Census 1845–40, 1850–45, 1860–50 (rural)         | link, maybe, unborn, not found |  | Description of specific nuances relating to 1845–1860 censuses. Description of census sources and introduction to guiding principles. Potential links are generated using a rule-based algorithm based on name and age.   |
| 2   | 16-06-2020; ALA 0.5.2               | Census 1845–40, 1850–45, 1860–50 (all areas)     | unlink                         | Links saved automatically. Possibility of searching using <b>herred</b> and <i>amt</i> . | Includes guidance for linking urban areas as well as rural.   |
| 3   | 28-09-2020; ALA 1.0.1, 1.1.0        | Census (as in 2) + 1880–60, 1901–1880, 1901–1885 | link_w_occupation              | .  | Introduction of a prioritisation order for linking criteria: household information takes precedence, followed by name, birth place, age, civil status, and occupation. Considerations for census distance, naming conventions, treatment of elderly individuals, and enumerator variations. |

Table 39 continued from previous page

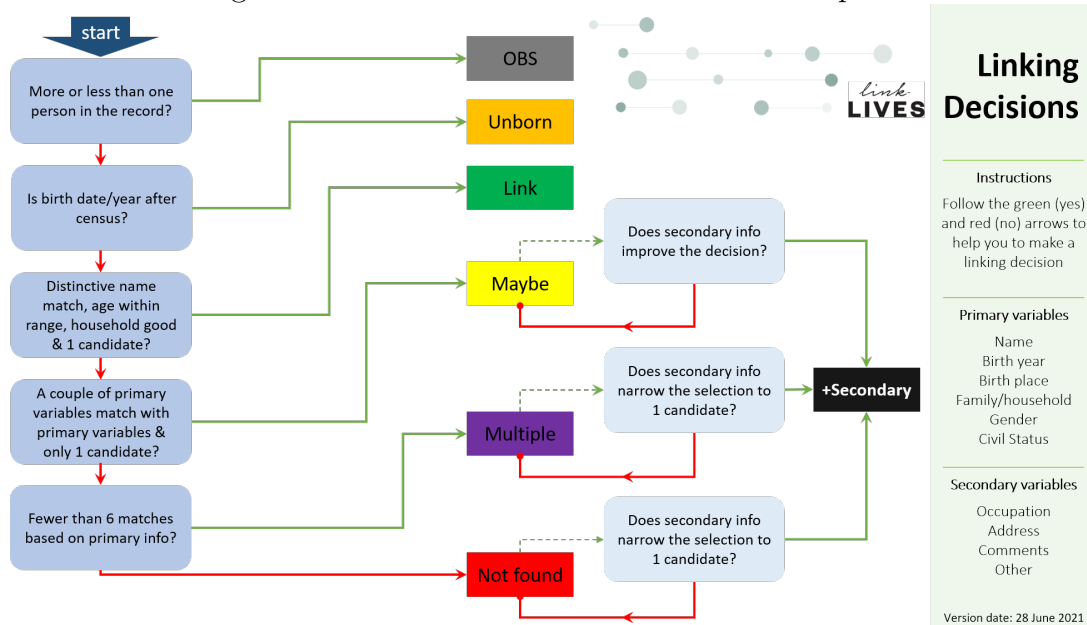
| Nr. | Date & ALA version         | Sources covered               | New link decisions   | New functions  | Recommendation changes  |
|-----|----------------------------|-------------------------------|--|--|---|
| 4   | 15-03-2021; ALA 1.38/9     | As in BP 3 (ALA 1.0.1, 1.1.0) | more_than_c (AKA: OBS), sec-ondary_may sec-ondary_not sec-ondary_mul | Use OBS if a record is unlinkable (which removes it from linkables total).   | Introduction of a “what’s new” feature. Launch of linking PR burials. Guidance on handling limited household context information. Linking of individual deaths in order, rather than in household structures. Introduction of a linking decision chart, providing a structured approach to making linking decisions. Potential links not available for PR burials. Guidance on primary–secondary variable use. Availability of name statistics to allow linkers to assess the rarity of given and surnames. |
| 5+6 | 01-07-2021; ALA 1.4, 1.5.1 | PR burials only               |  | Introduction of simultaneous two-people search function within a set range of person appearances. Any link decision confirmed by residence deemed a “+secondary one”. Logging of crash errors for faster debugging. All fields in the full info sections can appear as pop-out dialog boxes for ease of reading. | Introduction of linking by convenience. Defines how linkers are to select and link records within specific year ranges for a particular parish (PR burials only). Clarification of “multiple” = 16 plausible candidates, and not found = 5 plausible candidates. Guidance for handling non-standard records, e.g. two people in one record, duplicates, empty addresses. Name statistics suspended.   |
| 7   | 14-10-2021; ALA 1.6        | PR marriages and PR baptisms  |  | New “proximity” search field to narrow second-person search results to a limited distance from person 1.   | Introduction of marriage and baptism parish registers. Direction of linking forwards. Baptisms: use the parish of baptism as a proxy for birthplace, with the greatest weight, followed by the place of actual birth. If the baptism parish is empty, it is assumed the child was born there. No potential links available for PR.  |

Table 39 continued from previous page

| Nr. | Date & ALA version  | Sources covered                            | New link decisions | New functions   | Recommendation changes                               |
|-----|---------------------|--|--------------------|---|--|
| 8   | 30-11-2021; ALA 1.8 | All existing sources to 1845–1901 censuses |                    | An autosearch function replaces potential links files (potentials automatically generated based on name and age blocking). New search function for “dd-mm-yyyy searches”. | Introduction of new system to track linking process. |

One of the most important changes was the introduction of two separate decisions for cases where both primary and secondary information were used, which led to creating the decision tree in figure 39, which was used ever since.

Figure 39: Decision tree aid for our domain experts.



## 10.7 Definition of disagreements

To compare the input of two linkers on the same person appearance, we defined disagreements as follows:

- Censuses: a disagreement between two decisions was identified if:
  - If one linker linked a person appearance and the other one did not
  - They both linked a person appearance from origin source to target source but the id chosen was different. That means that if they found the same person appearance in the target source but one of them chose “link” and the other “maybe”, this inconsistency was not substantial enough to be sent for review.
  - The cases where unborn, not found and missing created an **id**= -1 so any disagreement of those was not sent either for further review.
- Copenhagen Burials and Parish records: after the introduction of primary and secondary decision, the possibilities of disagreement increased substantially. For instance, both linkers could find the same person appearance in the target source but linker 1 selected a “maybe” while linker 2 selected a “multiple+secondary”, that is they assessed “multiple” based on primary variable but upgraded the decision to “maybe” after having a look at secondary variables. We assigned a specific code to each possible combination

and established rules for what was considered a disagreement. As our main focus was to preserve the possibility of using the resulting dataset with information on whether primary or secondary information was used, that was the key issue to distinguish what was considered a disagreement that required manual evaluation. For cases, like that above, “maybe” vs “multiple+secondary”, there was an underlying disagreement on the first decision, positive vs negative, so it was sent to review. In table 40 we display the table used for deciding whether a combination of decisions was considered a disagreement that required sending it back to review. The last column of the table also shows what value we assigned to those combinations that we did not consider disagreements that required being sent for review.

Table 40: Possible combinations of linker decisions and actions taken about them.

| Decision code | Decision combinations              | Sent to review? | Conservative decision |
|---------------|------------------------------------|-----------------|-----------------------|
| 00a           | 00a-same-link                      | No              |                       |
| 00b           | 00b-different-link                 | Yes             |                       |
| 01a           | 01a-same-link-maybe                | No              | maybe                 |
| 01b           | 01b-different-link-maybe           | Yes             |                       |
| 02            | 02-link-not_found                  | Yes             |                       |
| 03            | 03-link-unborn                     | Yes             |                       |
| 04            | 04-link-multiple                   | Yes             |                       |
| 05a           | 05a-same-link-maybe_secondary      | No              | maybe_secondary       |
| 05b           | 05b-different-link-maybe_secondary | Yes             |                       |
| 05            | 05-link-maybe_secondary            | Yes             |                       |
| 06a           | 06a-same link-link occup           | No              | link occup            |
| 06a           | 06a-same-link-not_found_secondary  | No              | not_found_secondary   |
| 06b           | 06b-different link-link occup      | Yes             |                       |



Table 40 continued from previous page

| Decision code | Decision combinations                   | Sent to review? | Conservative decision |
|---------------|---|-----------------|-----------------------|
| 06b           | 06b-different-link-not_found_secondary  | Yes             |                       |
| 07a           | 07a-same-link-multiple_secondary        | No              | multiple_secondary    |
| 07b           | 07b-different-link-multiple_secondary   | Yes             |                       |
| 08            | 08-link-more_than_one                   | Yes             |                       |
| 0x            | 0x-link-forgotten                       | Yes             |                       |
| 10a           | 10a-same-maybe-link                     | No              | maybe                 |
| 10b           | 10b-different-maybe-link                | Yes             |                       |
| 11a           | 11a-same-maybe                          | No              |                       |
| 11b           | 11b-different-maybe                     | Yes             |                       |
| 12            | 12-maybe-not_found                      | Yes             |                       |
| 13            | 13-maybe-unborn                         | Yes             |                       |
| 14            | 14-maybe-multiple                       | Yes             |                       |
| 15a           | 15a-same-maybe-maybe_secondary          | No              | maybe_secondary       |
| 15b           | 15b-different-maybe-maybe_secondary     | Yes             |                       |
| 15            | 15-maybe-maybe_secondary                | Yes             |                       |
| 16a           | 16a-same-maybe-not_found_secondary      | No              | not_found_secondary   |
| 16b           | 16b-different-maybe-link_occup          | Yes             |                       |
| 16b           | 16b-different-maybe-not_found_secondary | Yes             |                       |

Table 40 continued from previous page

| Decision code | Decision combinations                  | Sent to review? | Conservative decision |
|---------------|--|-----------------|-----------------------|
| 17a           | 17a-same-maybe-multiple_secondary      | No              | multiple_secondary    |
| 17b           | 17b-different-maybe-multiple_secondary | Yes             |                       |
| 18            | 18-maybe-more_than_one                 | Yes             |                       |
| 1x            | 1x-maybe-forgotten                     | Yes             |                       |
| 20            | 20-not_found-link                      | Yes             |                       |
| 21            | 21-not_found-maybe                     | Yes             |                       |
| 22            | 22-not_found-not_found                 | No              |                       |
| 23            | 23-not_found-unborn                    | No              | unborn                |
| 24            | 24-not_found-multiple                  | No              | not_found             |
| 25            | 25-not_found-maybe_secondary           | Yes             |                       |
| 26            | 26-not_found-not_found_secondary       | Yes             |                       |
| 26            | 26-not_found-not_found_secondary       | Yes             |                       |
| 27            | 27-not_found-multiple_secondary        | Yes             |                       |
| 28            | 28-not_found-more_than_one             | Yes             |                       |
| 2x            | 2x-not_found-forgotten                 | Yes             |                       |
| 30            | 30-unborn-link                         | Yes             |                       |
| 31            | 31-unborn-maybe                        | Yes             |                       |
| 32            | 32-unborn-not_found                    | No              | unborn                |

Table 40 continued from previous page

| Decision code | Decision combinations              | Sent to review? | Conservative decision |
|---------------|------------------------------------|-----------------|-----------------------|
| 33            | 33-unborn-unborn                   | No              |                       |
| 34            | 34-unborn-multiple                 | No              | unborn                |
| 35            | 35-unborn-maybe_secondary          | Yes             |                       |
| 36            | 36-unborn-not_found_secondary      | Yes             |                       |
| 3x            | 3x-unborn-forgotten                | No              | unborn                |
| 40            | 40-multiple-link                   | Yes             |                       |
| 41            | 41-multiple-maybe                  | Yes             |                       |
| 42            | 42-multiple-not_found              | No              | not_found             |
| 43            | 43-multiple-unborn                 | No              | unborn                |
| 44            | 44-multiple-multiple               | No              |                       |
| 45            | 45-multiple-maybe_secondary        | Yes             |                       |
| 46            | 46-multiple-link<br>occup          | Yes             |                       |
| 46            | 46-multiple-not_found_secondary    | Yes             |                       |
| 47            | 47-multiple-multiple_secondary     | Yes             |                       |
| 4x            | 4x-multiple-forgotten              | Yes             |                       |
| 50a           | 50a-same-maybe_secondary-link      | No              | maybe_secondary       |
| 50b           | 50b-different-maybe_secondary-link | Yes             |                       |

Table 40 continued from previous page

| Decision code | Decision combinations                                     | Sent to review? | Conservative decision |
|---------------|---|-----------------|-----------------------|
| 50            | 50-<br>maybe_secondary-<br>link                           | Yes             |                       |
| 51a           | 51a-same-<br>maybe_secondary-<br>maybe                    | No              | maybe_secondary       |
| 51b           | 51b-different-<br>maybe_secondary-<br>maybe               | Yes             |                       |
| 52            | 52-<br>maybe_secondary-<br>not_found                      | Yes             |                       |
| 53            | 53-<br>maybe_secondary-<br>unborn                         | Yes             |                       |
| 54            | 54-<br>maybe_secondary-<br>multiple                       | Yes             |                       |
| 55a           | 55a-same-<br>maybe_secondary                              | No              |                       |
| 55b           | 55b-different-<br>maybe_secondary                         | Yes             |                       |
| 56a           | 56a-same-<br>maybe_secondary-<br>not_found_secondary      | No              | not_found_secondary   |
| 56b           | 56b-different-<br>maybe_secondary-<br>not_found_secondary | Yes             |                       |
| 57a           | 57a-same-<br>maybe_secondary-<br>multiple_secondary       | No              | multiple_secondary    |
| 58            | 58-<br>maybe_secondary-<br>more_than_one                  | Yes             |                       |

Table 40 continued from previous page

| Decision code | Decision combinations                        | Sent to review? | Conservative decision |
|---------------|--|-----------------|-----------------------|
| 5x            | 5x-maybe_secondary-forgotten                 | Yes             |                       |
| 60a           | 60a-same link occup-link                     | No              | link occup            |
| 60a           | 60a-same-not_found_secondary-link            | No              | not_found_secondary   |
| 60b           | 60b-different-not_found_secondary-link       | Yes             |                       |
| 61a           | 61a-same link occup-maybe                    | Yes             |                       |
| 61a           | 61a-same-not_found_secondary-maybe           | No              | not_found_secondary   |
| 61b           | 61b-different-not_found_secondary-maybe      | Yes             |                       |
| 62            | 62-link occup-not_found                      | Yes             |                       |
| 62            | 62-not_found_secondary-not_found             | Yes             |                       |
| 63            | 63-not_found_secondary-unborn                | Yes             |                       |
| 64            | 64-link occup-multiple                       | Yes             |                       |
| 64            | 64-not_found_secondary-multiple              | Yes             |                       |
| 65a           | 65a-same-not_found_secondary-maybe_secondary | No              | not_found_secondary   |

Table 40 continued from previous page

| Decision code | Decision combinations                                | Sent to review? | Conservative decision |
|---------------|--|-----------------|-----------------------|
| 65b           | 65b-different-not_found_secondary-maybe_secondary    | Yes             |                       |
| 66a           | 66a-same-not_found_secondary                         | No              |                       |
| 66b           | 66b-different-not_found_secondary                    | Yes             |                       |
| 66a           | 66a-same link occup                                  | No              |                       |
| 67a           | 67a-same-not_found_secondary-multiple_secondary      | No              | not_found_secondary   |
| 67b           | 67b-different-not_found_secondary-multiple_secondary | Yes             |                       |
| 68            | 68-not_found_secondary-more_than_one                 | Yes             |                       |
| 6x            | 6x-not_found_secondary-forgotten                     | Yes             |                       |
| 70a           | 70a-same-multiple_secondary-link                     | No              | multiple_secondary    |
| 70b           | 70b-different-multiple_secondary-link                | Yes             |                       |
| 71a           | 71a-same-multiple_secondary-maybe                    | No              | multiple_secondary    |
| 71b           | 71b-different-multiple_secondary-maybe               | Yes             |                       |
| 72            | 72-multiple_secondary-not_found                      | Yes             |                       |

Table 40 continued from previous page

| Decision code | Decision combinations                            | Sent to review? | Conservative decision |
|---------------|--|-----------------|-----------------------|
| 74            | 74-multiple_secondary-multiple                   | Yes             |                       |
| 75a           | 75a-same-multiple_secondary-maybe_secondary      | No              | multiple_secondary    |
| 75b           | 75b-different-multiple_secondary-maybe_secondary | Yes             |                       |
| 76a           | 76a-same-multiple_secondary-not_found_secondary  | No              | not_found_secondary   |
| 77a           | 77a-same-multiple_secondary                      | No              |                       |
| 77b           | 77b-different-multiple_secondary                 | Yes             |                       |
| 7x            | 7x-multiple_secondary-forgotten                  | Yes             |                       |
| 80            | 80-more_than_one-link                            | Yes             |                       |
| 81            | 81-more_than_one-maybe                           | Yes             |                       |
| 82            | 82-more_than_one-not_found                       | Yes             |                       |
| 85            | 85-more_than_one-maybe_secondary                 | Yes             |                       |
| 87            | 87-more_than_one-multiple_secondary              | Yes             |                       |
| 88            | 88-more_than_one-more_than_one                   | No              |                       |
| 8x            | 8x-more_than_one-forgotten                       | Yes             |                       |
| x0            | x0-link-forgotten                                | Yes             |                       |

Table 40 continued from previous page

| Decision code | Decision combinations            | Sent to review? | Conservative decision |
|---------------|----------------------------------|-----------------|-----------------------|
| x1            | x1-maybe-forgotten               | Yes             |                       |
| x2            | x2-not_found-forgotten           | Yes             |                       |
| x3            | x3-unborn-forgotten              | No              | unborn                |
| x5            | x5-maybe_secondary-forgotten     | Yes             |                       |
| x6            | x6-not_found_secondary-forgotten | Yes             |                       |
| x7            | x7-multiple_secondary-forgotten  | Yes             |                       |
| x8            | x8-more_than_one-forgotten       | Yes             |                       |

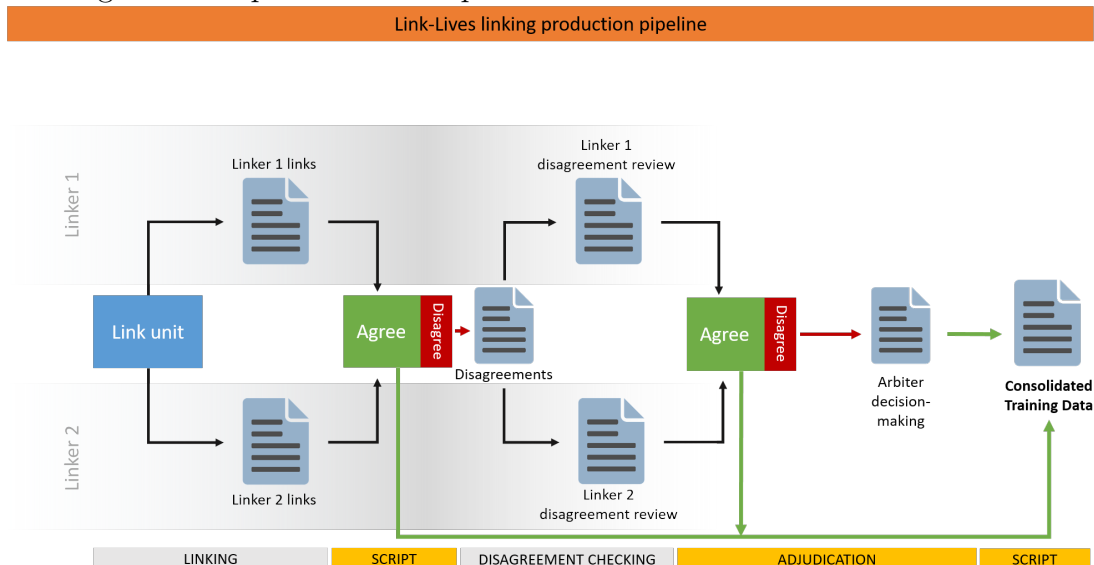
## 10.8 Pipeline from links files to benchmark dataset

The actual implementation of the computer-assisted domain expert linking described in section 6.1.4 involved many persons, platforms and formats. In figure 40 we describe the overall process and below we describe the actual step by step process involved in it:

1. Two linkers linked all person apperances of an assigned linking unit using the linking interface, ALA. Their links were saved automatically in a `links.csv` file which the program produced. They pasted a copy of their latest links file to our Trello administration board for the unit in question and labeled it complete.
2. Periodically a linking coordinator looked at the administration board and when a unit had two complete `links.csv` files from two assigned linkers, she ran a Python script to compare the decisions. The script generated an Excel file of all decisions where there was a disagreement between linker 1 and linker 2's decisions. (See definitions in section 10.7).
3. Each linker saved this file locally and worked through the disagreements, providing a decision and explaining it briefly in a short text. The options for decisions at this stage were:



Figure 40: Pipeline from outputs of a linker to the benchmark dataset.



- “linker1”: if they, on reassessment, agreed with Linker 1 (independently on whether it was the linker herself or not)
  - “linker2”: if they, on reassessment, agreed with Linker 2
  - “none”: if they on reassessment decided on a decision that was neither linker 1’s nor linker 2’s.
4. These two files were returned to the linking coordinator via the Trello board process.
  5. The coordinator merged both Linker 1 and Linker 2’s disagreement review files. This action divided the results in two types:
    - Moderate disagreement: if both linkers agreed after review. The records were not examined further.
    - Enduring disagreement: if there were still disagreements
  6. The file with enduring disagreements was sent to a third independent linker, who acted as an arbiter. On occasion, the linking coordinator performed the role of arbiter. This decision was based on the text description of the linkers and sometimes by attempting the link themselves to examine the possibilities through the linking software. This file was then passed back to the linking coordinator via the Trello board. We have included a document summarizing guidelines for arbitration as part of the **Link-Lives Paradata** accompanying documentation<sup>40</sup>.
  7. The coordinator then created a consolidated file for each linking unit using a script relevant to that source type.

40. The terminology used in the working documents is slightly different than the one presented in the guide as the project members used shortcuts to refer to them practically before we could describe a common framework. In the documentation, disagreements are called conflicts and arbiters are called conflict solvers.

- For all link decisions that were not deemed disagreements but were still not identical, the script downgraded the decision to the more conservative one according to the decisions listed in 40).
  - The script accounted for variable name differences between different source types, a range of different link decisions according to each source type, human error in typing anything other than what was prescribed in the adjudicated file and identified when there were anomalies, e.g. if a link had been forgotten
  - Finally, several fixed variables were added to aid traceability: date of creating a consolidated version from a linking unit, linking software version and linkers' initials.
8. Once consolidated data files were created for all units of a given source type were complete, they were amalgamated into one large dataset for that source type and then into a combined dataset for all sources. As this occurred over a long period of time, the original scripts, hard-coded to the specific sources and files, were replaced by automations that required less and less human intervention and ensure replicability.

## 10.9 Training linkers

### 10.9.1 Selection and training of linkers

As described in section 6.1.4, we chose as “linkers” persons with a background in History and had some prior interest/training in historical demography and 19th century Danish social history within the context of the project. This allowed us to ensure that our linkers had a minimal background knowledge of the sources and historical context and the project.

We involved a fairly large amount of persons and they were associated to the project in a variety of ways: from senior researcher to interns. We mostly had hired student assistants, research assistants and attracted six interns. Interns were students from the University of Copenhagen during their Master in History, who chose to carry out an academic internship with the project at either the National Archives or the University of Copenhagen for 15 or 30 ECTS credits (half or a full semester). They were unpaid positions but students received academic credit compensation.<sup>41</sup> Moreover, during the Coronavirus pandemic of 2020 we were also able to recruit a voluntary genealogist and four archivists during some months, which also had a sufficient historical background.<sup>42</sup>

Most of the data was linked during a very intensive period, rich in resources, from February 2020 to April 2022. From April 2022 until September 2023, it continued at a minimal pace. We started with censuses, followed by Copenhagen burials, and then parish records (burials, marriages and baptisms). In total more

---

41. For interns, linking was a core part of their tasks but only for a few hours a day, the rest of the time they were involved in other tasks in the project so they could gain knowledge in all steps of data production.

42. The benchmark dataset would have been impossible without them and all their names are written in the first page with all project collaborators

than 30 linkers participated in the process. However, only 10 have been responsible for most of the work. We have had between 5-10 linkers linking simultaneously but never full time, only doing a maximum of 2-3 hours a day. We entrusted the role of arbiter only to linkers with a lot of experience. However, the role was not exclusive, so several domain experts could be performing arbitrating tasks at the same time.

### 10.9.2 Linking school

As described in section 6.1.4, in order to ensure a consistent approach to linking within the “Best Practices”, we developed a training course, loosely inspired by the short account given Bailey et al Bailey et al. (2017) of the work at Brigham University Linking Unit. The training course consisted on an initial live session, followed for online/physical meetings, a variety of materials on linking (including a guide, best practices, the software guide, videos, etc) and access to a live chat. New linkers were to link three specific linking units (usually parishes) after which their decisions were compared to a consolidated data file of the parish that the project already had produced. They then received files with disagreements they need to review but they were not explicitly told what they were being compared to and their answers were only reviewed to get a sense of how they understood the process. This allowed us to mimic the process that they would follow as part of the regular process, which allowed for better understanding and improved consistency. Overall, by design, we limited trainee or domain expert linking activity (linking or revising files) to a maximum of three hours a day to ensure sufficient concentration on the task and allow adequate breaks. While we never policed it, we made sure to transmit the message that quality was always more important than quantity. We include an example of the documentation provided to new linkers in the project as part of the document **Link-Lives Paradata**.

### 10.9.3 First linking school design

We developed the first Linking School session as we developed and consolidated the process of linking the first rural census areas 1845-1860. Given that this coincided with the first lockdown of Coronavirus pandemic in 2020, we had access to some additional employees at the archive so there were 8 new recruits. This first session was done completely online.

### 10.9.4 Linking school sessions

Over time of construction of the Benchmark dataset there were 4 iterations of linking school, that coincided with times where new team members joined the team (mostly driven by interns). At each of the times, we slightly reformulated the format to better fit the profiles and numbers of new recruits (see table 41). Given the advances of the project, we needed to provide updated documentation to the new team members (ALA guide, best practices) as well as select units for linking that were relevant for the needs of the project.

The latest Linking School was set up as a two-week training program, with formal in-person presentations followed by practical linking exercises and a review

Table 41: Linking School Sessions held.

| No | Start date | Linker no. | Linker type                         | Sources                                  | Years                                    | No. linked | Comments   |
|----|------------|------------|-------------------------------------|--|--|------------|--|
| 1  | April 2020 | 8          | Historians                          | Rural census–census                      | 1850 to 1845; 1860 to 1850               | 946        | During lock-down. Ad hoc email instructions.                 |
| 2  | July 2020  | 7          | 5 historians; 1 intern; 1 volunteer | Rural census–census; Urban sample census | 1850 to 1845; 1860 to 1850; 1901 to 1880 | 976        | Standard email instructions.                                 |
| 3  | Aug 2021   | 3          | 2 interns; 1 historian              | PR burials–census                        | 1845 to 1850; 1850 to 1860; 1860 to 1880 | 678        | Standard email instructions and regular Link-Ins             |
| 4  | Feb 2022   | 4          | 3 interns; 1 historian              | PR burials–census; PR baptisms–census    | 1860 to 1880; 1845 to 1850; 1850 to 1860 | 631        | Formal delivery of training, monitoring and ongoing Link-Ins |

process, followed by weekly Link-Ins for ongoing support. There is a clear impact of this slow buildup of competences, especially with review of their links after each unit and the feedback.

## 10.10 Linking continuing monitoring: workshops and the “link-in”

In order to maintain consistency and not to deviate individually, we put in place check-ins of the linker group from time to time. During lockdown and until September 2021, we carried out a number of ad hoc online workshops, mostly on the occasion of the introduction of new ALA versions (and new sources). Afterwards, we created “Link-Ins” events, that were carried out both in person and online usually every month, with a set format to support group linking and update on software features or Best Practices.

“Link-Ins” were workshop-style linking sessions in which all linkers were asked to link a given sample either before the session or during it. It was an opportunity for everyone to discuss what they would do, and this helped to build consistency. A senior linker ran the session and gave a final decision on links. They were also responsible for recording any ambiguous cases which needed to be addressed in the Best Practices which were updated regularly.

We include an example of the slides of a Linkers workshop carried out in the earlier phases of the project to illustrate the type of uses we encountered as part of the document **Link-Lives Paradata**.

## 11 Benchmark dataset 1787-1840

This benchmark dataset, for the censuses 1787-1845 was created through the same method described in section 6.1 but with a slightly different implementation than that of the period 1845-1945, described in section 10. In this section, we only highlight the elements that differentiate either the approach or the implementation:

- The main difference with the benchmark dataset 1845-19021 was that these censuses were linked only for testing purposes, instead for both training and testing.
- The varying intervals between censuses were in general larger (1845-1840 is the shortest but all others are longer, 1840-1834, 1834-1801, 1801-1787) and the variables available for linking did not include information on birthplaces (with the exception of 1845).
- The linking units were selected by drawing a number of full pages from 21 parishes, from rural, urban and Copenhagen areas. Each census pairing consisted of approx. 600 linkable records.
- The newest version of ALA was used to carry out this linking set (v.1.9 and its associated User Guide and Best Practices) which meant it was possible to use the newest linking features in ALA (see section 39 for Best Practices describing ALA features and recommendations)

Overall we linked data from 21 distinct linking units that correspond to 20 geographies (which were parishes, streets/parishes in towns or sample years in Copenhagen burials), corresponding to 2,538 person appearances. (See full list in table 56 in the appendix).

## 12 Codebooks

### 12.1 Codebook for censuses





Table 42 continued: Codebook for censuses.

[illegible]

Table 42 continued: Codebook for censuses.

| Variable name          | Type            | Description  | Orig.  | Lev.          | 1787 | 1801 | 1834 | 1840 | 1845 | 1850 | 1860 | 1880 | 1885 | 1901 |
|------------------------|-----------------|--|--------|---------------|------|------|------|------|------|------|------|------|------|------|
| nr_ægteskab            | num             | Number of marriages for each individual.   | source | indiv.        | x    | x    |      |      |      |      |      |      |      |      |
| Erhverv                | string          | Content of either occupation (and also position in the household in some years) (See section 4.1.2). | source | indiv.        | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    |
| Kommentar              | string          | Comments to the source/transcription by the transcriber.   | DDD    | indiv.        | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    |
| Stilling_i_husstanden  | string          | Position in the house-hold (only used some years).   | source | indiv.        | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    |
| Antal_fam-<br>lier_hus | num             | Number of families in each household.  | source | house<br>hold | x    | x    | x    | x    | x    | x    |      | x    |      |      |
| handicaps              | string/<br>cat* | Handicaps/disabilities.  | source | indiv.        |      | x    |      |      |      | x    | x    | x    |      | x    |
| fødested               | string          | Place of birth   | source | indiv.        |      |      |      |      | x    | x    | x    | x    | x    | x    |
| trossamfund            | string          | Religion/faith.  | source | indiv.        |      |      |      |      |      |      | x    | x    | x    | x    |

Table 42 continued: Codebook for censuses.

| Variable name      | Type   | Description  | Orig.  | Lev.   | 1787 | 1801 | 1834 | 1840 | 1845 | 1850 | 1860 | 1880 | 1885 | 1901 |
|--------------------|--------|--|--------|--------|------|------|------|------|------|------|------|------|------|------|
| køn                | string | Sex. Inferred by a proportion of volunteers until explicitly recorded in 1870                | source | indiv. | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    |
| midlertid_oph_sted | string | Temporary whereabouts.   | source | indiv. | x    | x    | x    |      | x    | x    |      | x    |      |      |
| Adresse            | string | Address  | source | indiv. |      |      |      |      |      |      |      | x    | x    | x    |
| Matrikel           | string | Cadaster information   | source | indiv. |      |      |      |      |      |      |      | x    |      | x    |
| gadenr             | string | Street number. Often accidentally mixed with address or cadaster nr.                         | source | indiv. |      |      |      |      |      |      | x    |      |      | x    |
| etage              | string | Floor of the building  | source | indiv. |      |      |      |      |      |      |      |      |      | x    |
| forhus             | string | Whether the household lived in the front house on the address. (Often mixed with the floor). | source | indiv. |      |      |      |      |      |      | x    |      |      | x    |

Table 42 continued: Codebook for censuses.

[illegible]

Table 42 continued: Codebook for censuses.

[illegible]

Table 42 continued: Codebook for censuses.

[illegible]

Table 42 continued: Codebook for censuses.

[illegible]

Table 42 continued: Codebook for censuses.

| Variable name | Type   | Description   | Orig. | Lev.        | 1787 | 1801 | 1834 | 1840 | 1845 | 1850 | 1860 | 1880 | 1885 | 1901 |
|---------------|--------|---|-------|-------------|------|------|------|------|------|------|------|------|------|------|
| Sogn          | string | Name of parish or other geographic unit in <b>Type</b> .  | DDD   | transc unit | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    |
| Type          | string | Type of geographical unit (parish, town, borough etc.) whose name is described in <b>Sogn</b> . | DDD   | transc unit | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    |
| Rigsdel       | string | Country/territory (i.e. Denmark, Slesvig etc.).   | DDD   | transc unit | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    |
| Amt           | string | County.   | DDD   | transc unit | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    |
| Herred        | string | District.   | DDD   | transc unit | x    | x    | x    | x    | x    | x    | x    | x    | x    | x    |

“Orig.” stands for “Origin”, “Lev” for “Level”. We differentiate between origin=“source” as information captured from the source and “DDD”, processing on original information performed by DDD.  
transc unit = transcription unit.



## **12.2 Codebook for parish registers**

Table 43: Codebook for Parish Registers.

| Variable name              | Type   | Description   | Origin                     | Level      |
|----------------------------|--------|---|----------------------------|------------|
| <b>pa_id</b>               | num    | Person appearance id. The id is unique within a source, not across multiple sources. No missing values  | Link-Lives                 | individual |
| <b>event_id</b>            | num    | Ancestry's transcriptions are organised so that each row represents an event, and each event is assigned an ID number. No missing values  | Ancestry                   | event      |
| <b>ImageFileName</b>       | string | Ancestry's name of image file used for transcription.No missing values  | Ancestry                   | page       |
| <b>ImageFolder</b>         | string | Ancestry's name for the roll the image and record come from. No missing values.   | Ancestry.No missing values | page       |
| <b>unique_identifier</b>   | num    | A unique ID number assigned to that specific record. No missing values.   | Ancestry                   | event      |
| <b>SequenceNumber</b>      | num    | Original record sequence per page in the format ####, where the middle two digits indicate the order on the page, the first digit is always a 1 and the final digit is always a 0. The first entry on a page is always 10.No missing values | Ancestry                   | event      |
| <b>SourceCollectionID</b>  | num    | Collection identification number or similar from the archive.   | Ancestry                   | event      |
| <b>SourceDescription</b>   | string | Description from the archive.   | Ancestry                   | source     |
| <b>SourceReferenceNumb</b> | string | Reference number or call number.  | Ancestry                   | source     |
| <b>SourceComments</b>      | string | Comments written by transcribers.   | Ancestry                   | event      |

Table 43 continued: Codebook for Parish Registers.

| Variable name          | Type   | Description   | Origin   | Level      |
|------------------------|--------|---|----------|------------|
| <b>SourceYearRange</b> | string | Year range of the dataset.  | Ancestry | event      |
| <b>BrowseLevel</b>     | string | Used to identify browse levels in a collection. It represents the county ( <i>amt</i> ).No missing values.      | Ancestry | event      |
| <b>BrowseLevel1</b>    | string | Used to identify browse levels in a collection. It represents the parish ( <i>sogn</i> ). No missing values.    | Ancestry | event      |
| <b>BrowseLevel2</b>    | string | Used to identify browse levels in a collection. It represents the year range. No missing values.                | Ancestry | event      |
| <b>KeyedParish</b>     | string | Parish keyed from image.  | Ancestry | event      |
| <b>Notes</b>           | string | Notes about the transcription, e.g. “udøbt barn” (unbaptized child) for burials with no given names.            | Ancestry | event      |
| <b>NamePrefix</b>      | string | Name(s) classified in transcription as prefixes (e.g. titles).  | Mapped   | individual |
| <b>GivenName</b>       | string | Name(s) classified in transcription as given first and middle name(s).  | Mapped   | individual |
| <b>Surname</b>         | string | Name(s) classified in transcription as surnames.  | Mapped   | individual |
| <b>NameSuffix</b>      | string | Name(s) classified in transcription as suffixes (eg. “junior” and “senior”).                                    | Mapped   | individual |
| <b>GivenNameAlias</b>  | string | Alternative first and middle name(s).   | Mapped   | individual |
| <b>SurnameAlias</b>    | string | Alternative surname(s) for an individual.   | Mapped   | individual |
| <b>MaidenName</b>      | string | Name(s) classified in transcription as maiden name(s) ( i.e. a woman’s surname(s) at birth or before marriage). | Mapped   | individual |

Table 43 continued: Codebook for Parish Registers.

| Variable name     | Type   | Description  | Origin | Level      |
|-------------------|--------|--|--------|------------|
| <b>Gender</b>     | string | Gender of main person in event is derived from division into men and women in the parish register. When events are split into person appearances, parents are assigned a gender according to their role, i.e. mothers = female, and fathers = male. Specifically for Marriages: the main person of a marriage event will be the groom (male unless otherwise stated). When events are split into person appearances, the spouse is assigned the opposite gender to the main person, unless SpouseGender is filled in the original event level transcription. “Kvinde” = female, “Mandlige” = male. | Mapped | individual |
| <b>BirthDay</b>   | string | Gregorian day of the month in which the person was born. For spouses content is transferred from SpouseBirthYear at event level.   | Mapped | individual |
| <b>BirthMonth</b> | string | The month in which the person was born, written in letters or abbreviated. For spouses content is transferred from SpouseBirthMonth at event level.  | Mapped | individual |
| <b>BirthYear</b>  | string | The year in which the person was born. For spouses content is transferred from SpouseBirthYear at event level.   | Mapped | individual |

Table 43 continued: Codebook for Parish Registers.

| Variable name            | Type   | Description   | Origin   | Level      |
|--------------------------|--------|---|----------|------------|
| <b>BirthPlace</b>        | string | Used when geographic areas given as main person's place of birth are not parsed into separate city, county, state, or country fields. For spouses, this variable is a direct transfer of all content from SpouseBirthPlace when events are split into <b>pa.ids</b> . | Mapped   | individual |
| <b>BirthParish</b>       | string | Name of parish where the main person of the event was born.   | Ancestry | event      |
| <b>BirthMunicipality</b> | string | Name of municipality where the main person of the event was born.   | Ancestry | event      |
| <b>BirthState</b>        | string | Name of state where the main person of the event was born.  | Ancestry | event      |
| <b>BaptismAge</b>        | string | Age at the time of the event for an individual. Field contains instances of “ <i>dødfødt</i> ” (stillbirth) as well as a number of ages over 1, suggesting this field contains data if the child was not a (live) infant.   | Ancestry | event      |
| <b>BaptismDay</b>        | string | Gregorian day of the month on which the main person was baptized. Occasionally registered as the name of a season/holiday/saint's day.  | Ancestry | event      |
| <b>BaptismMonth</b>      | string | The month in which the main person was baptized, written in letters or abbreviated.   | Ancestry | event      |
| <b>BaptismYear</b>       | string | The year the main person was baptized.  | Ancestry | event      |
| <b>BaptismPlace</b>      | string | Used when geographic areas are not parsed into separate city, county, state, or country fields.   | Ancestry | event      |

Table 43 continued: Codebook for Parish Registers.

| Variable name               | Type   | Description   | Origin   | Level |
|-----------------------------|--------|---|----------|-------|
| <b>BaptismParish</b>        | string | Name of parish.   | Ancestry | event |
| <b>BaptismMunicipality</b>  | string | Name of municipality.   | Ancestry | event |
| <b>BaptismCounty</b>        | string | Name of county.   | Ancestry | event |
| <b>BaptismState</b>         | string | Name of state.  | Ancestry | event |
| <b>BaptismCountry</b>       | string | Name of country.  | Ancestry | event |
| <b>ConfirmationDay</b>      | string | Gregorian day of the month on which an event took place. Often registered as the name of a season (mostly spring or autumn)/holiday/church calendar date etc. | Ancestry | event |
| <b>ConfirmationMonth</b>    | string | The month in which an event took place, written in letters or abbreviated.  | Ancestry | event |
| <b>ConfirmationYear</b>     | string | The year an event took place.   | Ancestry | event |
| <b>ConfirmationPlace</b>    | string | Used when geographic areas are not parsed into separate city, county, state, or country fields.   | Ancestry | event |
| <b>ConfirmationParish</b>   | string | Name of parish.   | Ancestry | event |
| <b>ConfirmationMunicipa</b> | string | Name of municipality.   | Ancestry | event |
| <b>ConfirmationCounty</b>   | string | Name of county.   | Ancestry | event |
| <b>ConfirmationState</b>    | string | Name of state.  | Ancestry | event |
| <b>ConfirmationCountry</b>  | string | Name of country.  | Ancestry | event |
| <b>ArrivalDay</b>           | string | Gregorian day of the month on which an event took place. Occasionally registered as the name of a season/ holiday/saint's day.                                | Ancestry | event |

Table 43 continued: Codebook for Parish Registers.

| Variable name              | Type   | Description   | Origin   | Level |
|----------------------------|--------|---|----------|-------|
| <b>ArrivalMonth</b>        | string | The month in which an event took place, written in letters or abbreviated.  | Ancestry | event |
| <b>ArrivalYear</b>         | string | The year an event took place. Sometimes written as a range of years.  | Ancestry | event |
| <b>ArrivalAge</b>          | string | Age in years of main person at the time an event took place.  | Ancestry | event |
| <b>ArrivalPlace</b>        | string | Used when geographic areas are not parsed into separate city, county, state, or country fields.                               | Ancestry | event |
| <b>ArrivalParish</b>       | string | Name of parish.   | Ancestry | event |
| <b>ArrivalMunicipality</b> | string | Name of municipality.   | Ancestry | event |
| <b>ArrivalCounty</b>       | string | Name of county.   | Ancestry | event |
| <b>ArrivalState</b>        | string | Name of state.  | Ancestry | event |
| <b>ArrivalCountry</b>      | string | Name of country.  | Ancestry | event |
| <b>DepartureDay</b>        | string | Gregorian day of the month on which an event took place. Occasionally registered as the name of a season/holiday/saint's day. | Ancestry | event |
| <b>DepartureMonth</b>      | string | The month in which an event took place, written in letters or abbreviated.  | Ancestry | event |
| <b>DepartureYear</b>       | string | The year an event took place. Sometimes written as a range of years.  | Ancestry | event |
| <b>DepartureAge</b>        | string | Age in years of main person at the time an event took place.  | Ancestry | event |

Table 43 continued: Codebook for Parish Registers.

| Variable name                | Type   | Description  | Origin   | Level |
|------------------------------|--------|--|----------|-------|
| <b>DeparturePlace</b>        | string | Used when geographic areas are not parsed into separate city, county, state, or country fields.                                | Ancestry |       |
| <b>DepartureParish</b>       | string | Name of parish.  | Ancestry | event |
| <b>DepartureMunicipality</b> | string | Name of municipality.  | Ancestry | event |
| <b>DepartureCounty</b>       | string | Name of county.  | Ancestry | event |
| <b>DepartureState</b>        | string | Name of state.   | Ancestry | event |
| <b>DepartureCountry</b>      | string | Name of country.   | Ancestry | event |
| <b>MarriageDay</b>           | string | Gregorian day of the month on which an event took place. Occasionally registered as the name of a season/ holiday/saint's day. | Ancestry | event |
| <b>MarriageMonth</b>         | string | The month in which an event took place, written in letters or abbreviated.   | Ancestry | event |
| <b>MarriageYear</b>          | string | The year an event took place.  | Ancestry | event |
| <b>MarriageAge</b>           | string | Age in years of individual at the time an event took place.  | Ancestry | event |
| <b>MarriagePlace</b>         | string | Used when geographic areas are not parsed into separate city, county, state, or country fields.                                | Ancestry | event |
| <b>MarriageParish</b>        | string | Name of parish.  | Ancestry | event |
| <b>MarriageMunicipality</b>  | string | Name of municipality.  | Ancestry | event |
| <b>MarriageCounty</b>        | string | Name of county.  | Ancestry | event |
| <b>MarriageState</b>         | string | Name of state.   | Ancestry | event |



Table 43 continued: Codebook for Parish Registers.

| Variable name            | Type   | Description   | Origin   | Level |
|--------------------------|--------|---|----------|-------|
| <b>MarriageCountry</b>   | string | Name of country.  | Ancestry | event |
| <b>DeathDay</b>          | string | Gregorian day of the month of death (of main person).   | Ancestry | event |
| <b>DeathMonth</b>        | string | Month of death (of main person) written in letters or abbreviated.  | Ancestry | event |
| <b>DeathYear</b>         | string | Year of death (of main person).   | Ancestry | event |
| <b>DeathAge</b>          | string | Age in years at death. Can also be expressed as “ <i>dødfødt</i> ”, / “ <i>Dødfødt</i> ”, “ <i>todtgeboren</i> ” for stillbirths or number of months (format (x)x/12) for infants. In Baptisms it is used for stillborns. | Ancestry | event |
| <b>DeathPlace</b>        | string | Used when geographic areas are not parsed into separate city, county, state, or country fields.   | Ancestry | event |
| <b>DeathParish</b>       | string | Name of parish.   | Ancestry | event |
| <b>DeathMunicipality</b> | string | Name of municipality.   | Ancestry | event |
| <b>DeathState</b>        | string | Name of state.  | Ancestry | event |
| <b>BurialDay</b>         | string | Gregorian day of the month on which an event took place. Occasionally registered as the name of a season/ holiday/saint’s day.  | Ancestry | event |
| <b>BurialMonth</b>       | string | The month in which an event took place, written in letters or abbreviated.  | Ancestry | event |
| <b>BurialYear</b>        | string | Year of burial (main person).   | Ancestry | event |

Table 43 continued: Codebook for Parish Registers.

| Variable name                | Type   | Description   | Origin   | Level |
|------------------------------|--------|---|----------|-------|
| <b>BurialAge</b>             | string | Age in years of main person at the time they were buried. Can also be expressed as “ <i>dødfødt</i> ”/ “ <i>Dødfødt</i> ” for stillbirths or number of months (format (x)x/12) for infants. In Baptisms it is only used for stillborns. | Ancestry | event |
| <b>BurialPlace</b>           | string | Used when geographic areas are not parsed into separate city, county, state, or country fields.   | Ancestry | event |
| <b>BurialParish</b>          | string | Name of parish.   | Ancestry | event |
| <b>BurialMunicipality</b>    | string | Name of municipality.   | Ancestry | event |
| <b>BurialCounty</b>          | string | Name of county.   | Ancestry | event |
| <b>BurialState</b>           | string | Name of state.  | Ancestry | event |
| <b>BurialCountry</b>         | string | Name of country.  | Ancestry | event |
| <b>ResidenceAge</b>          | string |   | Ancestry | event |
| <b>ResidenceParish</b>       | string | Name of parish where the person(s) of the event resides.  | Ancestry | event |
| <b>ResidenceMunicipality</b> | string | Name of municipality where the person(s) of the event resides.  | Ancestry | event |
| <b>ResidenceCounty</b>       | string | Name of county where the person(s) of the event resides.  | Ancestry | event |
| <b>ReligiousDay</b>          | string | (empty variable)  | Ancestry | event |
| <b>ReligiousMonth</b>        | string | (empty variable)  | Ancestry | event |
| <b>ReligiousYear</b>         | string | (empty variable)  | Ancestry | event |

Table 43 continued: Codebook for Parish Registers.

| Variable name                | Type   | Description   | Origin   | Level |
|------------------------------|--------|---|----------|-------|
| <b>ReligiousAge</b>          | string | (empty variable)  | Ancestry | event |
| <b>ReligiousParish</b>       | string | (empty variable)  | Ancestry | event |
| <b>ReligiousMunicipality</b> | string | (empty variable)  | Ancestry | event |
| <b>ReligiousState</b>        | string | (empty variable)  | Ancestry | event |
| <b>VitalDay</b>              | string | Gregorian day of the month on which an event took place. Occasionally registered as the name of a season/holiday/saint's day. | Ancestry | event |
| <b>VitalMonth</b>            | string | The month of the event, written in letters or abbreviated.  | Ancestry | event |
| <b>VitalYear</b>             | string | The year of the event. Year ranges occur.   | Ancestry | event |
| <b>VitalAge</b>              | string | (Main person's) Age in years at the time of the event. (Baptism: used for stillbirths).                                       | Ancestry | event |
| <b>VitalPlace</b>            | string | Used when geographic areas given as place of event are not parsed into separate city, county, state, or country fields.       | Ancestry | event |
| <b>VitalParish</b>           | string | Name of parish where the event took place.  | Ancestry | event |
| <b>VitalMunicipality</b>     | string | Name of the municipality where the event took place.  | Ancestry | event |
| <b>VitalState</b>            | string | Name of the state where the event took place.   | Ancestry | event |
| <b>VitalCountry</b>          | num    | Name of the country where the event took place.   | Ancestry | event |
| <b>VitalCounty</b>           | string | Name of the county where the event took place.  | Ancestry | event |
| <b>VitalTownship</b>         | string | Name of the township where the event took place.  | Ancestry | event |

Table 43 continued: Codebook for Parish Registers.

| Variable name              | Type   | Description   | Origin   | Level |
|----------------------------|--------|---|----------|-------|
| <b>ArrivalCity</b>         | string | Name of city.   | Ancestry | event |
| <b>BaptismCity</b>         | string | Name of city.   | Ancestry | event |
| <b>BirthCountry</b>        | string | Name of city.   | Ancestry | event |
| <b>BurialCity</b>          | string | Name of city.   | Ancestry | event |
| <b>ConfirmationAge</b>     | string | Age at the time of the event for an individual.   | Ancestry | event |
| <b>ConfirmationCity</b>    | string | Name of city.   | Ancestry | event |
| <b>DepartureCity</b>       | string | Name of city.   | Ancestry | event |
| <b>MarriageCity</b>        | string | Name of city.   | Ancestry | event |
| <b>VitalCity</b>           | string | Name of the city where the event took place.  | Ancestry | event |
| <b>FatherGivenName</b>     | string | First and middle name(s) of person identified as father of main person in event.              | Ancestry | event |
| <b>FatherSurname</b>       | string | Surname(s) of person identified as father of main person in event.                            | Ancestry | event |
| <b>FatherNameSuffix</b>    | string | Suffix(e) (eg. “junior” and “senior”) of person identified as father of main person in event. | Ancestry | event |
| <b>FatherGivenNameAlia</b> | string | Alternative first and middle name(s) of person identified as father of main person in event.  | Ancestry | event |
| <b>FatherSurnameAlias</b>  | string | Alternative surname(s) of person identified as father of main person in event.                | Ancestry | event |
| <b>MotherNamePrefix</b>    | string | Prefix(es) (eg. titles) of person identified as mother of main person in event.               | Ancestry | event |

Table 43 continued: Codebook for Parish Registers.

| Variable name              | Type   | Description   | Origin   | Level |
|----------------------------|--------|---|----------|-------|
| <b>MotherGivenName</b>     | string | First and middle name(s) of perso identified as mother of main person in event.   | Ancestry | event |
| <b>MotherSurname</b>       | string | Surname(s) for person identified as mother of main person in event.   | Ancestry | event |
| <b>MotherNameSuffix</b>    | string | Suffix(es) (eg. “junior” and “senior”) for person identified as mother of main person in event.                                 | Ancestry | event |
| <b>MotherMaidenName</b>    | string | Maiden name(s) ( ie. a woman’s surname(s) at birth or before marriage) for person identified as mother of main person in event. | Ancestry | event |
| <b>MotherGivenNameAli.</b> | string | Alternative first and middle name(s) for person identified as mother of main person in event.                                   | Ancestry | event |
| <b>MotherSurnameAlias</b>  | string | Alternative surname(s) for person identified as mother of main person in event.   | Ancestry | event |
| <b>MotherResidenceAge</b>  | string | Field can contain age data for the child’s mother.  | Ancestry | event |
| <b>SpouseNamePrefix</b>    | string | Prefix(es) (eg. titles) of person identified as spouse of main person in event. For marriages, “spouse” is the bride.           | Ancestry | event |
| <b>SpouseGivenName</b>     | string | First and middle name(s) of person identified as spouse of main person in event. For marriages, “spouse” is the bride.          | Ancestry | event |
| <b>SpouseSurname</b>       | string | Surname(s) of person identified as spouse of main person in event. For marriages, “spouse” is the bride.                        | Ancestry | event |

Table 43 continued: Codebook for Parish Registers.

| Variable name               | Type   | Description   | Origin   | Level |
|-----------------------------|--------|---|----------|-------|
| <b>SpouseNameSuffix</b>     | string | Suffix(es) (eg. “junior” and “senior”) of person identified spouse of main person in event. For marriages, “spouse” is the bride.                                     | Ancestry | event |
| <b>SpouseMaidenName</b>     | string | Maiden name(s) ( i.e. a woman’s surname(s) at birth or before marriage) of person identified as spouse of main person in event. For marriages, “spouse” is the bride. | Ancestry | event |
| <b>SpouseGivenNameAlias</b> | string | Alternative first and middle name(s) of person identified as spouse of main person in event. For marriages, “spouse” is the bride.                                    | Ancestry | event |
| <b>SpouseSurnameAlias</b>   | string | Alternative surname(s) of person identified as spouse of main person in event. For marriages, “spouse” is the bride.  | Ancestry | event |
| <b>SpouseGender</b>         | string | Predicted gender of spouse.   | Ancestry | event |
| <b>SpouseMarriageAge</b>    | string | Age of spouse at marriage in years.   | Ancestry | event |
| <b>SpouseBirthDay</b>       | string | Gregorian day of the month of birth of person identified as spouse of main person in event. For marriages, “spouse” is the bride.                                     | Ancestry | event |
| <b>SpouseBirthMonth</b>     | string | Month of birth for person identified as spouse of main person in event. For marriages, “spouse” is the bride.   | Ancestry | event |
| <b>SpouseBirthYear</b>      | string | Year of birth for person identified as spouse of main person in event. For marriages, “spouse” is the bride.  | Ancestry | event |

Table 43 continued: Codebook for Parish Registers.

| Variable name                | Type   | Description   | Origin   | Level |
|------------------------------|--------|---|----------|-------|
| <b>SpouseBirthPlace</b>      | string | Place of birth for person identified as spouse of main person in event. For marriages, “spouse” is the bride.   | Ancestry | event |
| <b>FatherInLawGivenName</b>  | string | First and middle name(s) of person identified as father in law of main person in event. For marriages, fatherinlaw is the bride’s father.             | Ancestry | event |
| <b>FatherInLawSurname</b>    | string | Surname(s) of person identified as father in law of main person in event. For marriages, fatherinlaw is the bride’s father.                           | Ancestry | event |
| <b>FatherInLawNameSuffix</b> | string | Suffix(es) of person identified as father in law of main person in event. For marriages, fatherinlaw is the bride’s father.                           | Ancestry | event |
| <b>FatherInLawGivenName2</b> | string | Alternative first and middle name(s) of person identified as father in law of main person in event. For marriages, fatherinlaw is the bride’s father. | Ancestry | event |
| <b>FatherInLawSurname2</b>   | string | Alternative surname(s) of person identified as father in law of main person in event. For marriages, fatherinlaw is the bride’s father.               | Ancestry | event |
| <b>MotherInLawNamePrefix</b> | string | Prefix(es) of person identified as mother in law of main person in event. For marriages, motherinlaw is the bride’s mother.                           | Ancestry | event |
| <b>MotherInLawGivenName</b>  | string | First and middle name(s) of person identified as mother in law of main person in event. For marriages, motherinlaw is the bride’s mother.             | Ancestry | event |

Table 43 continued: Codebook for Parish Registers.

| Variable name             | Type   | Description  | Origin   | Level      |
|---------------------------|--------|--|----------|------------|
| <b>MotherInLawSurname</b> | string | Surname(s) of person identified as mother in law of main person in event. For marriages, motherinlaw is the bride's mother.  | Ancestry | event      |
| <b>MotherInLawNameSul</b> | string | Suffix(es) of person identified as mother in law of main person in event. For marriages, motherinlaw is the bride's mother.  | Ancestry | event      |
| <b>MotherInLawMaidenN</b> | string | Maiden name(s) of person identified as mother in law of main person in event. For marriages, motherinlaw is the bride's mother.  | Ancestry | event      |
| <b>MotherInLawGivenNa</b> | string | Alternative first and middle name(s) of person identified as mother in law of main person in event. For marriages, motherinlaw is the bride's mother.  | Ancestry | event      |
| <b>MotherInLawSurname</b> | string | Alternative surname(s) of person identified as mother in law of main person in event. For marriages, motherinlaw is the bride's mother.  | Ancestry | event      |
| <b>image_id</b>           | num    | ID for the image.  | Ancestry | page       |
| <b>folder_id</b>          | num    | ID for the roll the image and record come from.  | Ancestry | source     |
| <b>image_appearance</b>   | num    | Running number for the records on an image.  | Ancestry | page       |
| <b>event_type</b>         | string | Type of event. Each dataset represents a different event type, and should only contain that type of event. Possible values: arrival, baptism, burial, confirmation, death, departure, marriage | LL       | event      |
| <b>role</b>               | string | Roles extracted from event variables.  | LL       | individual |



Table 43 continued: Codebook for Parish Registers.

| Variable name | Type | Description  | Origin | Level |
|---------------|------|--|--------|-------|
| event_persons | num  | Number of person appearances extracted from event. | LL     | event |

Notes: “Mapped” denotes an original value from Ancestry that has been mapped by LL to this value in the transformation from events into person appearances. See section 5.2. Most variables are described as string type, despite mainly containing numerical values, because they contain nonnumeric values

## **12.3 Codebook for Copenhagen Burial Register**

Table 44: Codebook for Copenhagen Burial Register.

| Variable name | Type   | Description  | Origin                    | Level      |
|---------------|--------|--|---------------------------|------------|
| pa_id         | num    | Person appearance ID, unique within a source.  | Link-Lives                | individual |
| id            | num    | Burial ID from the Copenhagen City Archives).  | Copenhagen Ci<br>Archives | individual |
| number        | num    | Running number.  | original                  | individual |
| firstnames    | string | All first names of the deceased person(s), except nicknames.   | original                  | individual |
| lastname      | string | Last name of the deceased person. For young children, the last name is often derived from the last of a parent as recorded in the register.  | original                  | individual |
| birthname     | string | Maiden name for married/widowed women.   | original                  | individual |
| ageYears      | num    | Number of years recorded in deceased person's age. For stillborns, abortions, and infants with an age recorded in months/weeks/days/hours that adds up to less than 1 year, <b>ageYears</b> is transcribed as 0. | original                  | individual |
| ageMonth      | num    | Number of months recorded in the deceased person's age.  | original                  | individual |
| ageWeeks      | num    | Number of weeks recorded in the deceased person's age.   | original                  | individual |

Table 44 continued: Codebook for Copenhagen Burial Register.

| Variable name    | Type               | Description   | Origin                    | Level      |
|------------------|--------------------|---|---------------------------|------------|
| ageDays          | num                | Number of days recorded in the deceased person's  | original                  | individual |
| ageHours         | num                | The number of hours recorded in the deceased per  | original                  | individual |
| dateOfBirth      | date<br>(dd-mm-yyy | Date of birth, recorded in the burial protocols from<br>january 1913.   | original                  | individual |
| dateOfDeath      | date<br>(dd-mm-yyy | The date of death.  | original                  | individual |
| yearOfBirth      | num                | Year of birth calculated from recorded age and year<br>of death.  | Copenhagen Ci<br>Archives | individual |
| deathplace       | string             | The place of death, or the place the body was found   | drop-down                 | individual |
| civilstatus      | string             | Marital status of the deceased. Also includes the status<br>children explicitly recorded as either illegitimate,<br>legitimate, orphans, or foundlings. | drop-down                 | individual |
| adressOutsideCph | string             | Address/place of residence of the deceased person(<br>outside of Copenhagen, Frederiksberg, or<br>Gentofte.   | original                  | individual |
| sex              | string             | Sex of the deceased, inferred from<br>the source  | interpreted               | individual |
| comment          | string             | Any extra information that does not fit in the other<br>columns, including transcribers comments.   | Volunteers                | individual |

Table 44 continued: Codebook for Copenhagen Burial Register.

| Variable name | Type   | Description   | Origin    | Level      |
|---------------|--------|---|-----------|------------|
| cemetery      | string | The cemetery or crematorium where the body was delivered to.  | drop-down | individual |
| chapel        | string | Place the body was stored before being moved.<br>If several places are recorded, they are in transcription separated by a comma.  | drop-down | individual |
| parish        | string | Parish registered in relation to the burial (common from ca. the 1880s).  | drop-down | individual |
| street        | string | Street name in the recorded place of residence of the deceased.   | drop-down | individual |
| hood          | string | Neighbourhood of the street in the deceased person's place of residence.  | drop-down | individual |
| street_unique | string | Unique street name. If a street name can refer to several different streets in the city, the neighbourhood is added (format: "[street name] ([neighbourhood])" to distinguish between them. | dropdown  | individual |
| street_number | num    | The numerical digits in the house number, if recorded as part of the address  | original  | individual |
| letter        | string | Letter in the house number  | original  | individual |
| floor         | string | Floor of the building.  | drop-down | individual |

Table 44 continued: Codebook for Copenhagen Burial Register.

| Variable name             | Type   | Description  | Origin    | Level      |
|---------------------------|--------|--|-----------|------------|
| institution               | string | Name of institution if not death place is not an ordinary address.   | drop-down | individual |
| institution_street        | string | Street name of the institution   | drop-down | individual |
| institution_hood          | string | Neighbourhood of the institution.  | drop-down | individual |
| institution_street_unique | string | The unique street name of the institution chosen from the list.  | drop-down | individual |
| institution_street_number | num    | The house number in the address of the institution chosen from the list.   | drop-down | individual |
| positions                 | string | Occupation or social position recorded for the deceased person or their close family (spouse, parent etc.)   | drop-down | individual |
| relationtypes             | string | The relation between the deceased person and each of the recorded occupations or positions, i.e. whether it is the current or former position of the deceased, their spouse, mother, father etc. | drop-down | individual |
| workplaces                | string | Place of occupation for the deceased person themselves or their close family members.  | drop-down | individual |
| deathcauses               | string | Cause of deaths. Multiple causes of death are separated by commas.   | drop-down | individual |

*drop-down = the variable is a transcription true to the wording of the original source, but through a partially standardised drop-down list.*

## **12.4 Codebook for LL harmonized data**



singlelinecheck=false, justification=centering

Table 45: Codebook for Link-Lives harmonized data.

| Variable name    | Type       | Description  | Origin     | Level      |
|------------------|------------|--|------------|------------|
| pa_id            | num        | Person appearance id. The ID is unique within a source.  | computed   | individual |
| name_cl          | harmonized | Lowercase full name after removing unwanted characters.  | cleaned    | individual |
| name             | string     | Standardised full name.  | harmonized | individual |
| first_names      | string     | Standardised names classified as first names.  | harmonized | individual |
| patronyms        | string     | Standardised names classified as patronyms.  | harmonized | individual |
| family_names     | string     | Standardised names classified as family names.   | harmonized | individual |
| maiden_names     | string     | Standardised names classified as maiden names.   | harmonized | individual |
| all_patronyms    | string     | All possible patronyms (standardised names). Includes constructed names based on husband/father names.     | harmonized | individual |
| all_family_names | string     | All possible family names (standardised names). Includes constructed names based on husband/father names.  | harmonized | individual |
| uncat_names      | string     | Unclassified standardised names.   | harmonized | individual |
| sex              | string     | The predicted sex.   | predicted  | individual |
| marital_status   | string     | Harmonized marital status.   | original   | individual |
| age              | num        | The age of the time of the event. Ages originally recorded in days or months has been calculated in years. | computed   | individual |

Table 45: Codebook for LL standardised data (continued).

| Variable name  | Type   | Description   | Origin     | Level      |
|----------------|--------|---|------------|------------|
| birth_date     | date   | Date expressed in format is yyyy-mm-dd.   | harmonized | individual |
| birth_year     | num    | Year of birth (from birth_date if it was recorded or calculated from age and event date). | computed   | individual |
| birth_month    | num    | Month of birth (if birthdate was recorded in the original source).                        | computed   | individual |
| birth_day      | num    | Day of birth (if birthdate was recorded in the original source).                          | computed   | individual |
| event_date     | date   | Date of event. The format is yyyy-mm-dd.  | computed   | event      |
| event_year     | num    | Year of event.  | computed   | event      |
| event_month    | num    | Month of event.   | computed   | event      |
| event_day      | num    | Day of the event.   | computed   | event      |
| birth_place_cl | string | Full cleaned birth place string.  | clean      | individual |
| birth_place    | string | Full standardised birth place string.   | harmonized | individual |
| birth_location | string | Standardised birth place information classified as location.                              | computed   | individual |
| birth_parish   | string | Standardised birth place information classified as parish ( <i>sogn</i> )                 | computed   | individual |
| birth_town     | string | Standardised birth place information classified as town ( <i>købstad</i> )                | harmonized | individual |
| birth_county   | string | Standardised birth place information classified as county ( <i>am</i> )                   | harmonized | individual |

Table 45: **Codebook for LL standardised data (continued).**

| Variable name       | Type   | Description  | Origin     | Level      |
|---------------------|--------|--|------------|------------|
| birth_foreign_place | string | Standardised birth place information classified as foreign_place, i.e. a reference to a city or region in another country. This category does not include country names. | harmonized | individual |
| birth_country       | string | Standardised birth place information classified as country.  | harmonized | individual |
| event_location      | string | Standardised event place information classified as location.   | harmonized | event      |
| event_parish        | string | Standardised event place information classified as parish ( <i>sogndal</i> ).  | harmonized | event      |
| event_district      | string | Standardised event place information classified as district ( <i>herred</i> ).   | harmonized | event      |
| event_town          | string | Standardised event place information classified as town ( <i>købstad</i> ).  | harmonized | event      |
| event_county        | string | Standardised event place information classified as county ( <i>amt</i> ).  | harmonized | event      |
| event_country       | string | Standardised event place information classified as country.  | harmonized | event      |
| household_id        | num    | Household ID. Uses multiple variables to get a better separation of households. Only relevant for censuses.  | computed   | household  |
| household_position  | string | Standardised household position. Only relevant for censuses.   | harmonized | individual |
| role                | string | Whether the person appearance is the main person in the event. Not relevant for censuses.  | computed   | individual |
| event_type          | string | Type of event (e.g. census, burial, baptism, etc.).  | computed   | event      |
| book_id             | num    | ID for the collection the image and record come from.  | computed   | collection |

Table 45: Codebook for LL standardised data (continued).

| Variable name    | Type   | Description  | Origin   | Level      |
|------------------|--------|--|----------|------------|
| image_id         | string | ID for the image when available                                | computed | page       |
| image_appearance | num    | Running number for the records on an image or in a collection. | computed | individual |
| source_id        | num    | ID for the source.   | computed | source     |

## **12.5 Codebook for censuses for ALA**

Table 46: Codebook for ALA census data.

| ALA variable        | Transcribed             | Harmonized | Other transformations/ Notes   |
|---------------------|-------------------------|------------|--|
| id                  |                         | pa_id      |  |
| Place               | Sogn,<br>Herred,<br>Amt |            | Concatenation of the listed variables. Used for selecting parish in startup menu |
| Res. parish         | Sogn                    |            |  |
| Res. county         | Amt                     |            |  |
| Res.<br>information | ejendom_navn            |            | Concatenation of the listed variables  |
|                     | Adresse                 |            |  |
|                     | gadenr                  |            |  |
|                     | forhus                  |            |  |
|                     | Matr_nr_adresse         |            |  |
|                     | Stednavn                |            |  |
|                     | Matrikel                |            |  |
| Occ.<br>information | Erhverv                 |            |  |
|                     | Stilling_i_husstander   |            |  |
|                     | Erhvervssted            |            |  |
|                     | Arbejdsplads            |            |  |
| Marital st.         | Civilstand              |            |  |
| Sex                 |                         | sex        |  |

Table 46 continued from previous page

| ALA variable | Transcribed          | Harmonized   | Other transformations/ Notes   |
|--------------|----------------------|--------------|--|
| Name         | navn                 |              |  |
| Birth place  | Stednavn<br>fødested |              | <b>fødested</b> used after 1845<br><b>Stednavn</b> used for censuses without recorded birthplace (1787, 1801, 1834, 1840). |
| Age          |                      | age          |  |
| Birth year   |                      | birth_year   |  |
| Birth month  |                      | birth_month  |  |
| Birth day    |                      | birth_day    |  |
| R.no.        | løbenr_i_indtastning |              |  |
| HH. id       |                      | household_id |  |
| Plot no.     | Matrikel             |              |  |

## **12.6 Codebook for parish registers for ALA**



Table 47: Codebook for ALA parish register data.

| ALA variable   | Source type  | Transcribed    | Harmonized | Other transformations/<br>Notes   |
|----------------|--------------|----------------|------------|---|
| ImageFolder    | PR baptisms  | ImageFolder    |            |   |
|                | PR marriages |                |            |   |
|                | PR burials   |                |            |   |
| ImageFileName  | PR baptisms  | ImageFileName  |            |   |
|                | PR marriages |                |            |   |
|                | PR burials   |                |            |   |
| id             | PR baptisms  | pa_id          |            |   |
|                | PR marriages |                |            |   |
|                | PR burials   |                |            |   |
| SequenceNumber | PR baptisms  | SequenceNumber |            |   |
|                | PR marriages |                |            |   |
|                | PR burials   |                |            |   |
| Page-seq.      | PR baptisms  | {page}-{Seq.}  |            | <b>page</b> is calculated from the first <b>image_id</b> in each <b>EventPlace</b> . <b>Seq.</b> indicates the order of persons appearing within each event, based on <b>role</b> and <b>image_appearance</b> . |
|                | PR marriages |                |            |   |
|                | PR burials   |                |            |   |

Table 47 continued from previous page

| ALA variable | Source type  | Transcribed                              | Harmonized | Other transformations/<br>Notes   |
|--------------|--------------|--|------------|---|
| Parish       | PR baptisms  | BrowseLevel                              |            |   |
|              | PR marriages | BrowseLevel1                             |            |   |
|              | PR burials   |  |            |   |
| Gender       | PR baptisms  |  |            |   |
|              | PR marriages | Gender                                   |            |   |
|              | PR burials   |  |            |   |
| Sex          | PR baptisms  |  |            |   |
|              | PR marriages |  | <i>sex</i> | Marriages: Where <b>role</b> is “spouse”, we assign opposite value of “Sex” |
|              | PR burials   |  |            |   |
| Role         | PR baptisms  |  |            |   |
|              | PR marriages | role                                     |            |   |
|              | PR burials   |  |            |   |
| Notes        | PR baptisms  |  |            |   |
|              | PR marriages | Notes                                    |            |   |
|              | PR burials   |  |            |   |
| Age          | PR baptisms  | Bapt: BaptismAge                         |            |   |
|              | PR marriages | Mar: MarriageAge<br>or SpouseMarriageAge |            |   |
|              | PR burials   | Bur: BurialAge                           |            |   |

Table 47 continued from previous page

| ALA variable | Source type                               | Transcribed   | Harmonized | Other transformations/<br>Notes   |
|--------------|---|---|------------|---|
| Full name    | PR baptisms<br>PR marriages<br>PR burials | Baptisms and marriages:<br>-NamePrefix<br>-GivenName<br>-Surname<br>-Alias<br>-NameSuffix<br>Notes<br>“f.” MaidenName |            | Concatenation of all available<br>listed variables in their role<br>specific permutations<br>(eg. <b>MotherGivenName</b> ). |
|              |   | Burials: DeathAge   |            |   |
|              |   | Notes   |            |   |
|              |   | -NamePrefix   |            |   |
|              |   | -GivenName  |            |   |
|              |   | -GivenNameAlias   |            |   |
|              |   | -Surname  |            |   |
|              |   | -SurnameAlias   |            |   |
|              |   | -Alias  |            |   |
|              |   | -NameSuffix   |            |   |
|              |   | “f.” MaidenName   |            |   |

Table 47 continued from previous page

| ALA variable        | Source type  | Transcribed    | Harmonized | Other transformations/<br>Notes |
|---------------------|--------------|----------------|------------|---------------------------------|
| Given name          | PR baptisms  | GivenName      |            |                                 |
|                     | PR marriages |                |            |                                 |
|                     | PR burials   |                |            |                                 |
| Given name<br>alias | PR baptisms  | GivenNameAlias |            |                                 |
|                     | PR marriages |                |            |                                 |
|                     | PR burials   |                |            |                                 |
| Maiden name         | PR baptisms  | MadenName      |            |                                 |
|                     | PR marriages |                |            |                                 |
|                     | PR burials   |                |            |                                 |
| Surname             | PR baptisms  | Surname        |            |                                 |
|                     | PR marriages |                |            |                                 |
|                     | PR burials   |                |            |                                 |
| Surname alias       | PR baptisms  | SurnameAlias   |            |                                 |
|                     | PR marriages |                |            |                                 |
|                     | PR burials   |                |            |                                 |
| Name prefix         | PR baptisms  | NamePrefix     |            |                                 |
|                     | PR marriages |                |            |                                 |
|                     | PR burials   |                |            |                                 |

Table 47 continued from previous page

| ALA variable | Source type  | Transcribed | Harmonized | Other transformations/<br>Notes |
|--------------|--------------|-------------|------------|---------------------------------|
| Name suffix  | PR baptisms  | NameSuffix  |            |                                 |
|              | PR marriages |             |            |                                 |
|              | PR burials   |             |            |                                 |
| Birth place  | PR baptisms  | BirthPlace  |            |                                 |
|              | PR marriages |             |            |                                 |
|              | PR burials   |             |            |                                 |
| Birth day    | PR baptisms  | BirthDay    |            |                                 |
|              | PR marriages |             |            |                                 |
|              | PR burials   |             |            |                                 |
| Birth month  | PR baptisms  | BirthMonth  |            |                                 |
|              | PR marriages |             |            |                                 |
|              | PR burials   |             |            |                                 |
| Birth year   | PR baptisms  | BirthYear   |            |                                 |
|              | PR marriages |             |            |                                 |
|              | PR burials   |             |            |                                 |

Table 47 continued from previous page

| ALA variable       | Source type                               | Transcribed            | Harmonized                            | Other transformations/<br>Notes  |
|--------------------|---|------------------------|---------------------------------------|--|
| <b>*Birth year</b> | PR baptisms<br>PR marriages<br>PR burials | Baptisms:<br>BirthYear | Marriages and bu<br><i>birth_year</i> | Baptisms: If role = Mother,<br><b>Birth year*</b> is calculated<br>by subtracting the standardised<br>age from standardised<br><b>event_year</b> . |
| <b>Event id</b>    | PR baptisms<br>PR marriages<br>PR burials | Event_id               |                                       |  |
| <b>Event years</b> | PR baptisms<br>PR marriages<br>PR burials | BrowseLevel2           |                                       |  |
| <b>Year min</b>    | PR baptisms<br>PR marriages<br>PR burials |                        |                                       | First year in "Event_year"<br>(BrowseLevel2).  |
| <b>Year max</b>    | PR baptisms<br>PR marriages<br>PR burials |                        |                                       | Last year in "Event years"<br>(BrowseLevel2).  |
| <b>Baptism day</b> | PR baptisms                               |                        | <i>event_day</i>                      |  |

Table 47 continued from previous page

| ALA variable       | Source type               | Transcribed                                    | Harmonized  | Other transformations/<br>Notes                                 |
|--------------------|---------------------------|--|---|---|
| Baptism month      | PR baptisms               |  | <i>event_month</i>                                |   |
| Baptism year       | PR baptisms               |  | <i>event_year</i>                                 |   |
| Baptism place      | PR baptisms               | BaptismParish<br>BaptismCounty<br>BaptismState |   | Concatenation of the listed variables,<br>where values present. |
| Baptism date       | PR baptisms               | BaptismDay<br>BaptismMonth-<br>BaptismYear     |   | Concatenation of the listed<br>variables, where populated.      |
| *Baptism date      | PR baptisms               |  | <i>event_day/<br/>event_month/<br/>event_year</i> | Concatenation of the listed<br>variables.                       |
| Death age          | PR baptisms<br>PR burials | DeathAge                                       |   |   |
| Burial age         | PR baptisms               | BurialAge                                      |   |   |
| Mother res.<br>age | PR baptisms               | MotherResidenceAge                             |   |   |
| Birth parish       | PR baptisms               | BirthParish                                    |   |   |

Table 47 continued from previous page

| ALA variable          | Source type  | Transcribed                                       | Harmonized  | Other transformations/<br>Notes                         |
|-----------------------|--------------|---|---|---|
| <b>*Birth date</b>    | PR baptisms  |   | <i>birth_day/</i><br><i>birth_month/</i><br><i>birth_year</i> | Concatenation of the listed variables.                  |
| <b>Birth date</b>     | PR baptisms  | BirthDay/BirthMonth/<br>BirthYear                 |   | Concatenation of the listed variables, where populated. |
| <b>Marriage day</b>   | PR marriages |   | <i>event_day</i>  |   |
| <b>Marriage month</b> | PR marriages |   | <i>event_month</i>  |   |
| <b>Marriage year</b>  | PR marriages |   | <i>event_year</i>   |   |
| <b>Marriage place</b> | PR marriages | MarriageParish<br>MarriageCounty<br>MarriageState |   | Concatenation of the listed variables, where populated. |
| <b>Marriage date</b>  | PR marriages | MarriageDay/<br>MarriageMonth/<br>MarriageYear    |   | Concatenation of the listed variables, where populated. |
| <b>*Marriage date</b> | PR marriages |   | <i>event_day/</i><br><i>event_month/</i><br><i>event_year</i> | Concatenation of the listed variables.                  |
| <b>Death place</b>    | PR burials   | DeathPlace  |   |   |



Table 47 continued from previous page

| ALA variable        | Source type | Transcribed                                    | Harmonized | Other transformations/<br>Notes   |
|---------------------|-------------|--|------------|---|
| <b>Res. place</b>   | PR burials  | ResidenceParish<br>ResidenceCounty             |            | Concatenation of the listed variables.  |
| <b>Burial place</b> | PR burials  | BaptismParish<br>BaptismCounty<br>BaptismState |            | Concatenation of the listed variables, where available.   |
| <b>Death day</b>    | PR burials  | DeathDay                                       |            |   |
| <b>Death month</b>  | PR burials  | DeathMonth                                     |            |   |
| <b>Death year</b>   | PR burials  | DeathYear                                      |            |   |
| <b>Burial date</b>  | PR burials  | BurialDay/<br>BurialMonth/<br>BurialYear       |            | Concatenation of the listed variables.  |
| <b>*Death date</b>  | PR burials  | DeathDay/<br>DeathMonth/<br>d_DeathYear        |            | To address the empty year cells and non-standard-year data, we reconstructed the year of death by generating a derived field d_DeathYear. |

Variables with “\*” were calculated combining existing transcribed or standardized variables and the “\*” is part of their name

## **12.7 Codebook for Copenhagen Burial Register for ALA**

Table 48: Codebook for ALA Copenhagen Burial Register data.

| ALA variable         | Transcribed                            | Harmonized        | Other transformations/ Notes      |
|----------------------|--|-------------------|-----------------------------------|
| <b>R.no.</b>         | number                                 |                   |                                   |
| <b>Date of death</b> | dateOfDeath                            |                   |                                   |
| <b>Death causes</b>  | deathcauses                            |                   |                                   |
| <b>Sex</b>           | sex                                    |                   | lightly cleaned                   |
| <b>Age</b>           | ageYears                               |                   | lightly cleaned                   |
| <b>Months</b>        | ageMonth                               |                   |                                   |
| <b>Weeks</b>         | ageWeeks                               |                   |                                   |
| <b>Days</b>          | ageDays                                |                   |                                   |
| <b>Hours</b>         | ageHours                               |                   |                                   |
| <b>Years</b>         | ageYears                               |                   |                                   |
| <b>Birth year</b>    |  | <i>birth_year</i> |                                   |
| <b>Name</b>          | firstnames<br>lastname                 |                   | concatenation of listed variables |
| <b>Birth name</b>    | birthname                              |                   |                                   |
| <b>Full name</b>     | firstnames<br>lastname<br>f. birthname |                   | concatenation of listed variables |

Table 48 continued from previous page

| ALA variable                | Transcribed               | Harmonized | Other transformations/ Notes   |
|-----------------------------|---------------------------|------------|--|
| <b>Res.<br/>information</b> | adressOutsideCph/         |            |  |
|                             | adressInCph               |            |  |
|                             | street                    |            | concatenation of listed variables  |
|                             | street_number             |            |  |
|                             | letter                    |            |  |
|                             | floor                     |            |  |
| <b>Res. parish</b>          | parish                    |            |  |
| <b>Marital st.</b>          | marital_status            |            |  |
| <b>Occ.<br/>information</b> |                           |            | Condition:IF relationtypes = “Eget erhverv”<br>or “forhenværende/pensioneret”,<br>Occ. information = position ELSE empty     |
| <b>Add.<br/>occupations</b> |                           |            | Condition:IF relationtypes $\neq$ “Eget erhverv”<br>or “forhenværende/pensioneret”, Add.<br>occupations=positions ELSE empty |
| <b>Institution</b>          | institution               |            |  |
|                             | institution_street        |            | concatenation of listed variables  |
|                             | institution_street_number |            |  |
| <b>Cemetery</b>             | cemetery                  |            |  |
| <b>Chapel</b>               | chapel                    |            |  |

Table 48 continued from previous page

| ALA variable | Transcribed | Harmonized | Other transformations/ Notes |
|--------------|-------------|------------|------------------------------|
| Comment      | comment     |            |                              |
| id           | pa_id       |            |                              |

## 12.8 Codebooks for links files

Table 49: Codebook for links from rule-based approach from release 1.

| Variable name   | Type | Description   |
|-----------------|------|---|
| link_id         | num  | Unique link identifier.   |
| source_id_o     | num  | ID of the source linked “from” (origin).  |
| source_id_t     | num  | ID of the source linked “to” (target).  |
| pa_id_o         | num  | Person appearance ID within the origin source. Identifies the person appearance in <b>source_id_o</b> . |
| pa_id_t         | num  | Person appearance ID within the target source. Identifies the person appearance in <b>source_id_t</b> . |
| method_id       | num  | Identifier of the method used to create the link. See <b>methods.csv</b> for details.                   |
| score           | num  | Score of the link produced by the rule-based models.  |
| iteration       | num  | Iteration number of the rule-based linking process (outer loop).  |
| iteration_inner | num  | Inner iteration number of the rule-based linking process.   |
| duplicates      | num  | Number of links that include the person appearance in the origin source.                                |

*Note: Structure of the file **links\_v1.2.csv**.*

Table 50: Codebook for links from machine learning approach from release 2.

| Variable name | Type | Description   |
|---------------|------|---|
| link_id       | num  | Unique link identifier.   |
| source_id_o   | num  | ID of the source linked <i>from</i> (origin).   |
| source_id_t   | num  | ID of the source linked <i>to</i> (target).   |
| pa_id_o       | num  | Person appearance ID within the origin source. Identifies the person appearance in <b>source_id_o</b> . |
| pa_id_t       | num  | Person appearance ID within the target source. Identifies the person appearance in <b>source_id_t</b> . |
| method_id     | num  | Identifier of the method used to create the link. See table 32.   |
| score         | num  | Score of the link produced by the machine learning models.  |
| score_diff    | num  | Difference between the score of this link and the next best potential person appearance match.          |
| in_lifecourse | num  | Indicator of whether a link_id has been included in lifecourses.  |

*Note: Structure of the file `links_v2.1.csv`.*

## 12.9 Codebook for life-courses

Table 51: Codebook for life-course files

| Variable name         | Type                                   | Description   |
|-----------------------|--|---|
| <b>life_course_id</b> | num                                    | Unique life-course identifier.  |
| <b>pa_ids</b>         | list of integers (comma “,” separated) | Person apperance ids making up the life-course. The i'th id notes <b>pa_id</b> within i'th <b>source_id</b> in sources. |
| <b>source_ids</b>     | list of integers (comma “,” separated) | The <b>source_ids</b> matching the <b>pa_ids</b> .  |
| <b>link_ids</b>       | list of integers (comma “,” separated) | Selection of <b>link_ids</b> used to create the life-course*  |
| <b>n_sources</b>      | num                                    | Number of sources in the life-course. Should equal the length of <b>pa_ids</b> and <b>source_ids</b> .                  |

\* While structurally the file is the same for both releases, the actual meaning and use of this variable is different. For release 1, given that the linking was sequential, each **link\_id** connects each of the combination of **source\_ids** and **pa\_ids**. However, for release 2, a given person appearance from parish registers could be linked to multiple censuses, so the same logic does not apply. For retrieving all **link\_ids** related to a given life-course, it is necessary to access the `links.csv` file.

See an example of how the data is used in section [7.3](#)



## 12.10 Codebook for synonym catalogues

We provide three synonym catalogues with release 2 for marital status, name and sex, that we have used to perform our harmonization. They are available in the files `SC_marital_status_v1.csv`, `SC_names_v1.csv` and `SC_sex_v1.csv` and their use is described in section 5.11.1 for marital status, section x for names and section 5.10.1 for sex.

They follow the same format, consisting of two columns:

- 'original': contains the original string after cleaning alphanumeric symbols, trimming spaces and conversion to lower case.
- 'standard': the standard we use for harmonization.

## **12.11 Codebook for Benchmark dataset 1787-1901**

Table 52: Codebook for benchmark dataset 1787-1901.

| Variable name   | Type   | Description  | Data level      |
|-----------------|--------|--|-----------------|
| source1_type    | string | Type of <b>source1</b> (census, parish registers, Copenhagen burials) (source 1 is origin source in <code>links.csv</code> file) | source          |
| source1         | num    | Year or year range of <b>source1</b> (source 1 is origin source in <code>links.csv</code> file)                                  | source          |
| source2         | num    | Year of <b>source2</b> census ( <b>source 2</b> is target source in <code>links.csv</code> file)                                 | source          |
| parish          | string | Name of linking unit (a parish, a street/neighbourhood name) or a year of burials)   | source          |
| id1             | num    | <b>pa_id</b> for in <b>source1</b>   | individual      |
| type            | string | Decision for each person appearance among valid options (e.g. “link”, “maybe”, etc.)   | individual      |
| id2             | num    | <b>pa_id</b> in <b>source2</b>   | individual      |
| linker1         | string | Linker1’s identity, anonymised to letters  | linking process |
| type_linker1    | string | Decision made by linker1   | individual      |
| id2_linker1     | num    | <b>pa_id</b> in <b>source2</b> for positive link decision by linker1   | individual      |
| linker2         | string | Linker2’s identity, anonymized to numbers  | linking process |
| type_linker2    | string | Decision made by linker2   | individual      |
| id2_linker2     | num    | <b>pa_id</b> in <b>source2</b> for positive link decision by linker2   | individual      |
| type_agreement  | string | Summary code for linker1 and linker2 type of agreement   | individual      |
| type_resolution | string | Extent to which the agreement was resolved (“uncontested”, “contested-moderate”, “contested-enduring”)                           | individual      |
| cs              | string | Arbiter identity, anonymised to letters  | linking process |
| timestamp       | date   | Date and time of the file generation   | linking process |
| ALA_version     | string | Version of ALA software  | linking process |

| Variable name | Type   | Description  | Data level      |
|---------------|--------|--|-----------------|
| period        | string | Period covered by benchmark dataset (1787–1845 or 1845–1901)   | linking process |
| type_method   | string | Method employed (“production” for Link-Lives approach, “experiment” for deviation with multiple linkers)         | linking process |
| index         | num    | Unique id for each record of the dataset   | individual      |
| linking_unit  | string | Full name of linking unit that acts as unique ID (concatenation of parish, <b>source1</b> , and <b>source2</b> ) | linking process |

## 13 Appendix

Table 53: Overview of datasets included in the release

| Folder            | Files                      |
|-------------------|----------------------------|
| main_datasets     | census_1787__v1__cl.csv    |
|                   | census_1787__v1__std.csv   |
|                   | census_1801__v1__cl.csv    |
|                   | census_1801__v1__std.csv   |
|                   | census_1834__v1__cl.csv    |
|                   | census_1834__v1__std.csv   |
|                   | census_1840__v1__cl.csv    |
|                   | census_1840__v1__std.csv   |
|                   | census_1845__v1__cl.csv    |
|                   | census_1845__v1__std.csv   |
|                   | census_1850__v1__cl.csv    |
|                   | census_1850__v1__std.csv   |
|                   | census_1860__v1__cl.csv    |
|                   | census_1860__v1__std.csv   |
|                   | census_1880__v1__cl.csv    |
|                   | census_1880__v1__std.csv   |
|                   | census_1885__v1__cl.csv    |
|                   | census_1885__v1__std.csv   |
|                   | census_1901__v1__cl.csv    |
|                   | census_1901__v1__std.csv   |
|                   | cbp_1860-1911__v1__std.csv |
| main_datasets/ALA | ALA_census_1787.csv        |
|                   | ALA_census_1801.csv        |
|                   | ALA_census_1834.csv        |
|                   | ALA_census_1840.csv        |
|                   | ALA_census_1845.csv        |

Table 53 continued from previous page

| Folder                | Files                                |
|-----------------------|--------------------------------------|
|                       | ALA_census_1850.csv                  |
|                       | ALA_census_1860.csv                  |
|                       | ALA_census_1880.csv                  |
|                       | ALA_census_1885.csv                  |
|                       | ALA_census_1901.csv                  |
|                       | ALA_copenhagen-burials_1850-1860.csv |
|                       | ALA_copenhagen-burials_1860-1880.csv |
|                       | ALA_copenhagen-burials_1880-1901.csv |
|                       | ALA_copenhagen-burials_1885-1901.csv |
|                       | ALA_copenhagen-burials_1901-2000.csv |
| links_and_lifecourses | benchmark_v1.xlsx                    |
|                       | life_courses_v1.2.csv                |
|                       | life_courses_v2.1.csv                |
|                       | links_v1.2.csv                       |
|                       | links_v2.csv                         |
| auxiliary             | SC_marital_status_v1.csv             |
|                       | SC_names_v1.csv                      |
|                       | SC_sex_v1.csv                        |

Table 54: Overview of main datasets that can be requested

| Institution                            | Files              |
|--|--------------------|
| Ancestry (through <i>Rigsarkivet</i> ) | baptism_v1_cl.csv  |
|  | baptism_v1_std.csv |

Table 54 continued from previous page

| Institution                           | Files                         |
|---------------------------------------|-------------------------------|
|                                       | marriage__v1__cl.csv          |
|                                       | marriage__v1__std.csv         |
|                                       | confirmation__v1__cl.csv      |
|                                       | confirmation__v1__std.csv     |
|                                       | burial__v1__cl.csv            |
|                                       | burial__v1__std.csv           |
|                                       | arrival__v1__cl.csv           |
|                                       | arrival__v1__std.csv          |
|                                       | departure__v1__cl.csv         |
|                                       | + 100 corresponding ALA files |
| <b><i>Københavns Stadsarkivet</i></b> | cbp_1860-1911__v1__cl.csv     |

Table 55: Number of person appearances in test dataset by origin linking unit (parish, street/neighbourhood or year) and target census for the period 1845-1901

| Parish                   | Census | Cph<br>Burials | PR<br>Bap-<br>tisms | PR<br>Burials | PR<br>Mar-<br>riages | Total |
|--------------------------|--------|----------------|---------------------|---------------|----------------------|-------|
| Aarhus,<br>Domsogn       | 0      | 0              | 0                   | 1,921         | 0                    | 1,921 |
| Aarre                    | 1,301  | 0              | 1,484               | 308           | 424                  | 3,517 |
| Aars                     | 400    | 0              | 0                   | 0             | 0                    | 400   |
| Almind (Brusk<br>Herred) | 587    | 0              | 0                   | 0             | 0                    | 587   |
| Alrø                     | 238    | 0              | 0                   | 0             | 0                    | 238   |
| Avernakø                 | 311    | 0              | 0                   | 0             | 0                    | 311   |
| Bringstrup               | 1,050  | 0              | 2,997               | 0             | 311                  | 4,358 |
| CBP1861                  | 0      | 200            | 0                   | 0             | 0                    | 200   |
| CBP1863                  | 0      | 186            | 0                   | 0             | 0                    | 186   |

Table 55 continued from previous page

| Parish                                    | Census | Cph<br>Burials | PR<br>Bap-<br>tisms | PR<br>Burials | PR<br>Mar-<br>riages | Total |
|---|--------|----------------|---------------------|---------------|----------------------|-------|
| CBP1866                                   | 0      | 165            | 0                   | 0             | 0                    | 165   |
| CBP1867                                   | 0      | 159            | 0                   | 0             | 0                    | 159   |
| CBP1870                                   | 0      | 159            | 0                   | 0             | 0                    | 159   |
| CBP1875                                   | 0      | 131            | 0                   | 0             | 0                    | 131   |
| CBP1879                                   | 0      | 135            | 0                   | 0             | 0                    | 135   |
| CBP1881                                   | 0      | 209            | 0                   | 0             | 0                    | 209   |
| CBP1885                                   | 0      | 211            | 0                   | 0             | 0                    | 211   |
| CBP1886                                   | 0      | 336            | 0                   | 0             | 0                    | 336   |
| CBP1887                                   | 0      | 151            | 0                   | 0             | 0                    | 151   |
| CBP1890                                   | 0      | 314            | 0                   | 0             | 0                    | 314   |
| CBP1895                                   | 0      | 166            | 0                   | 0             | 0                    | 166   |
| CBP1899                                   | 0      | 320            | 0                   | 0             | 0                    | 320   |
| Copenhagen,<br>Amager-<br>brokvarter      | 150    | 0              | 0                   | 0             | 0                    | 150   |
| Copenhagen,<br>Amagergade                 | 117    | 0              | 0                   | 0             | 0                    | 117   |
| Copenhagen,<br>Amaliegade                 | 111    | 0              | 0                   | 0             | 0                    | 111   |
| Copenhagen,<br>Christian-<br>shavnkvarter | 157    | 0              | 0                   | 0             | 0                    | 157   |
| Copenhagen,<br>Frederiksberg<br>sogn      | 0      | 0              | 255                 | 0             | 0                    | 255   |
| Copenhagen,<br>Frelser                    | 0      | 0              | 255                 | 0             | 0                    | 255   |
| Copenhagen,<br>FrimandKvarter             | 154    | 0              | 0                   | 0             | 0                    | 154   |
| Copenhagen,<br>Gothersgade                | 493    | 0              | 0                   | 0             | 0                    | 493   |



Table 55 continued from previous page

| Parish                                       | Census | Cph<br>Burials | PR<br>Bap-<br>tisms | PR<br>Burials | PR<br>Mar-<br>riages | Total |
|--|--------|----------------|---------------------|---------------|----------------------|-------|
| Copenhagen,<br>Helligaand                    | 0      | 0              | 261                 | 0             | 250                  | 511   |
| Copenhagen,<br>KlaedeboK-<br>varter          | 143    | 0              | 0                   | 0             | 0                    | 143   |
| Copenhagen,<br>Klaedebokvarter               | 146    | 0              | 0                   | 0             | 0                    | 146   |
| Copenhagen,<br>Koebmagerk-<br>varter         | 147    | 0              | 0                   | 0             | 0                    | 147   |
| Copenhagen,<br>Noerre kvarter                | 118    | 0              | 0                   | 0             | 0                    | 118   |
| Copenhagen,<br>Sankt Johannes                | 0      | 0              | 254                 | 0             | 250                  | 504   |
| Copenhagen,<br>Sankt Stefans                 | 0      | 0              | 0                   | 0             | 154                  | 154   |
| Copenhagen,<br>Sankt Trinitatis              | 0      | 0              | 0                   | 0             | 250                  | 250   |
| Copenhagen,<br>Sankta Annae<br>Vesterkvarter | 290    | 0              | 0                   | 0             | 0                    | 290   |
| Copenhagen,<br>Snarenskvarter                | 135    | 0              | 0                   | 0             | 0                    | 135   |
| Copenhagen,<br>Stormgade                     | 158    | 0              | 0                   | 0             | 0                    | 158   |
| Copenhagen,<br>Udenbysvesterk-<br>varter     | 130    | 0              | 0                   | 0             | 0                    | 130   |
| Copenhagen,<br>Vor Frue                      | 0      | 0              | 255                 | 0             | 254                  | 509   |
| Ebeltoft<br>Købstad                          | 995    | 0              | 0                   | 0             | 0                    | 995   |
| Elsborg                                      | 447    | 0              | 0                   | 0             | 0                    | 447   |

Table 55 continued from previous page

| Parish                                     | Census | Cph<br>Burials | PR<br>Bap-<br>tisms | PR<br>Burials | PR<br>Mar-<br>riages | Total |
|--|--------|----------------|---------------------|---------------|----------------------|-------|
| Fodslette<br>(Langelands<br>Søndre Herred) | 368    | 0              | 0                   | 0             | 0                    | 368   |
| Frederiks                                  | 439    | 0              | 0                   | 0             | 0                    | 439   |
| Gaerum                                     | 0      | 0              | 0                   | 0             | 201                  | 201   |
| Gilleleje                                  | 1,168  | 0              | 764                 | 0             | 699                  | 2,631 |
| Greve                                      | 599    | 0              | 0                   | 0             | 0                    | 599   |
| Gærum                                      | 1,381  | 0              | 3,426               | 434           | 282                  | 5,523 |
| Hallenslev                                 | 368    | 0              | 0                   | 0             | 0                    | 368   |
| Holsteinborg                               | 662    | 0              | 0                   | 0             | 0                    | 662   |
| Horreby                                    | 362    | 0              | 0                   | 0             | 0                    | 362   |
| Junget                                     | 1,117  | 0              | 2,549               | 434           | 390                  | 4,490 |
| Krønge                                     | 962    | 0              | 2,294               | 0             | 322                  | 3,578 |
| Kærum                                      | 2,497  | 0              | 549                 | 1,119         | 1,137                | 5,302 |
| Lønne                                      | 218    | 0              | 0                   | 0             | 0                    | 218   |
| Nebel                                      | 1,015  | 0              | 0                   | 0             | 0                    | 1,015 |
| Nyker                                      | 882    | 0              | 0                   | 0             | 0                    | 882   |
| Odense Købstad                             | 1,223  | 0              | 0                   | 0             | 0                    | 1,223 |
| Pedersker                                  | 768    | 0              | 0                   | 0             | 0                    | 768   |
| Rindum                                     | 340    | 0              | 0                   | 0             | 0                    | 340   |
| Rutsker                                    | 965    | 0              | 0                   | 0             | 0                    | 965   |
| Sejerø                                     | 1,030  | 0              | 0                   | 639           | 0                    | 1,669 |
| Selde                                      | 888    | 0              | 0                   | 644           | 0                    | 1,532 |
| Sevel                                      | 1,057  | 0              | 0                   | 0             | 0                    | 1,057 |
| Sir  | 395    | 0              | 0                   | 0             | 0                    | 395   |
| Skarrild                                   | 428    | 0              | 0                   | 0             | 0                    | 428   |
| Skeby                                      | 536    | 0              | 0                   | 0             | 0                    | 536   |

Table 55 continued from previous page

| Parish                           | Census | Cph<br>Burials | PR<br>Bap-<br>tisms | PR<br>Burials | PR<br>Mar-<br>riages | Total  |
|----------------------------------|--------|----------------|---------------------|---------------|----------------------|--------|
| Skoerring<br>(Framlev<br>Herred) | 301    | 0              | 0                   | 0             | 0                    | 301    |
| Skraem                           | 209    | 0              | 0                   | 0             | 0                    | 209    |
| Tjæreby (Stroe<br>Herred)        | 770    | 0              | 0                   | 0             | 0                    | 770    |
| Varpelev                         | 264    | 0              | 0                   | 0             | 0                    | 264    |
| Vester Egesborg                  | 437    | 0              | 0                   | 0             | 0                    | 437    |
| Vester Tostrup                   | 384    | 0              | 0                   | 0             | 0                    | 384    |
| Total                            | 29,811 | 2,842          | 15,343              | 5,499         | 4,924                | 58,419 |

Table 56: Number of person appearances in test dataset by origin linking unit (parish, street/neighbourhood or year) and target census for the period 1787-1845

| Parish                                  | Census | Total |
|---|--------|-------|
| Aarhus Købstad                          | 181    | 181   |
| Almind                                  | 71     | 71    |
| Bringstrup                              | 103    | 103   |
| Christiansø                             | 170    | 170   |
| Copenhagen,<br>Klaedebo Kvarter         | 168    | 168   |
| Copenhagen, Sankt<br>Annæ Oesterkvarter | 281    | 281   |
| Copenhagen, Snarens<br>Kvarter          | 142    | 142   |
| Ebeltoft Købstad                        | 65     | 65    |
| Faaborg Købstad                         | 149    | 149   |
| Fodslette                               | 90     | 90    |
| Gærum                                   | 177    | 177   |
| Hofetaten                               | 93     | 93    |

Table 56 continued from previous page

| Parish                    | Census | Total |
|---------------------------|--------|-------|
| Krønge                    | 156    | 156   |
| Kærum                     | 88     | 88    |
| Lønne                     | 172    | 172   |
| Neksø                     | 133    | 133   |
| Odense Købstad            | 72     | 72    |
| Selde                     | 101    | 101   |
| Skeby                     | 69     | 69    |
| Tjæreby (Stroe<br>Herred) | 57     | 57    |

## References

- Aagaard, Samantha Nordholt. 2024. “Ud og hjem igen?” PhD diss., University of Copenhagen.
- Abramitzky, Ran, Leah Platt Boustan, Katherine Eriksson, James J. Feigenbaum, and Santiago Pérez. 2019. *Automated Linking of Historical Data*. National Bureau of Economic Research. Accessed October 30, 2019. <http://www.nber.org/papers/w25825.pdf>.
- “Ændringer i forbindelse med nye EU-regler om datasikkerhed (GDPR).” Dansk Demografisk Database. 2018, June 1, 2018. <https://www.danishdemographicdatabase.org/da/ændringer-i-forbindelse-med-nye-eu-regler-om-datasikkerhed-gdpr>.
- Akgün, Özgür, Alan Dearle, Graham Kirby, Eilidh Garrett, Tom Dalton, Peter Christen, Chris Dibben, and Lee Williamson. 2020. “Linking Scottish vital event records using family groups.” *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 53, no. 2 (April 2, 2020): 130–146. ISSN: 0161-5440, accessed July 7, 2020. <https://doi.org/10.1080/01615440.2019.1571466>. <https://doi.org/10.1080/01615440.2019.1571466>.
- Antonie, Luiza, Kris Inwood, Daniel J. Lizotte, and J. Andrew Ross. 2014. “Tracking people over time in 19th century Canada for longitudinal analysis.” *Machine Learning* 95, no. 1 (April 1, 2014): 129–146. ISSN: 1573-0565, accessed April 15, 2021. <https://doi.org/10.1007/s10994-013-5421-0>. <https://doi.org/10.1007/s10994-013-5421-0>.
- Archives, Copenhagen City, and Villads Christensen (rådstuearkivar). 2024. *Begravelsesprotokoller for København 1861-1903*. Copenhagen City Archives Repository, April 29, 2024. <https://doi.org/10.71687/bvid.ksa.83>.
- Bailey, Martha, Connor Cole, Morgan Henderson, and Catherine Massey. 2017. *How Well Do Automated Methods Perform in Historical Samples? Evidence from New Ground Truth*. Working Paper 24019. National Bureau of Economic Research, November. Accessed July 2, 2018. <https://doi.org/10.3386/w24019>. <http://www.nber.org/papers/w24019>.
- Befolkningsforholdene i Danmark i det 19. Aarhundrede. Statistisk Tabelværk*. 1905. 5. rk., Litra A, nr. 5. København: Statens Statistiske Bureau.
- Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost: A Scalable Tree Boosting System.” *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.
- Christen, Peter. 2012. *Data matching, concepts and techniques for record linkage, entity resolution, and duplicate detection*. (online): Springer.
- Clausen, Nanna Floor. 2015. “The Danish Demographic Database—Principles and Methods for Cleaning and Standardisation of Data.” In *Population Reconstruction*, edited by Gerrit Bloothoof, Peter Christen, Kees Mandemakers, and Marijn Schraagen, 3–22. (online): Springer.

- Clausen, Nanna Floor, and Hans Jørgen Marker. 2000. "The Danish Data Archive." In *Handbook of International Historical Microdata for Population Research*, edited by Patricia Kelly Hall, Robert McCaa, and Gunnar Thorvaldsen, 79–92. Minneapolis: Minnesota Population Centre.
- Degn, Ole. 1991. *Alle skrives i mandtal, folketællinger og deres brug, Arkivernes informationsserie*. Rigsarkivet.
- Feigenbaum, James J. 2016. "A Machine Learning Approach to Census Record Linking." <https://jamesfeigenbaum.github.io/research/census-link-ml/>.
- Helgertz, Jonas, Joseph Price, Jacob Wellington, Kelly J. Thompson, Steven Ruggles, and Catherine A. Fitch. 2022. "A New Strategy for Linking U.S. Historical Censuses: A Case Study for the IPUMS Multigenerational Longitudinal Panel." *Historical Methods* 55 (1): 12–29. ISSN: 0161-5440. <https://doi.org/10.1080/01615440.2021.1985027>.
- Hertz, Michael. 1970. "De slesvigske og holstenske folketællinger 1803-1860." *Arkiv. Tidsskrift for arkivforskning udgivet af Rigsarkivet* III:49–61.
- Hjorth-Moritzsen, Line. 2026. "Mapping women's work in nineteenth-century Denmark – a digital history project." PhD diss., University of Copenhagen.
- Hovedresultaterne af Folketællingen i Kongeriget Danmark den 1ste Februar 1890, med tilhørende Befolkningskaart*. 1894. Statistisk Tabelværk. Rk. 4, Litra A ; 8.a. Det Statistiske Bureau / Danmarks Statistik.
- Johansen, Hans Christian. 1975. *Befolkningsudvikling og familiestruktur i det 18. århundrede*. Odense.
- . 2002. *Danish Population History*. Odense: University Press of Southern Denmark.
- Kællerød, Lars-Jakob Harding. 2019. "Adam Gottlob Øhlenschläger Hauch & Jeppe Smed Jensen: Et studie af etableringen af det efternavnetypologiske mellemnavn i Danmark i 1800-tallet." PhD diss., April.
- Kællerød, Lars-Jakob Harding, and Bárbara Ana Revuelta-Eugercios. 2015. "Identifying middle names in onomastic profiles – Exploring the usage of middle names in 19th century Denmark through the census of 1880." *Onoma* 50 (73). <https://doi.org/DOI:10.34158/ONOMA.50/2015/3>.
- Kahneman, Daniel. 2013. *Thinking, fast and slow* /. 1. paperback ed. New York: Farrar, Straus / Giroux. ISBN: 978-0-374-53355-7.
- "Kirkebøger." 1991. In *Dansk kulturhistorisk opslagsværk*, redacted by Erik Alstrup and Poul Erik Olsen, 440–442. Copenhagen: Dansk Historisk Fællesforening.
- "Kirkebøgerne." 1922. In *Salmonsens Konversations Leksikon*, redacted by Chr. Blangstrup, vol. XIII, 931. Copenhagen: Schultz Forlagsboghandel.
- Løkke, Anne. 1998. *Døden i barndommen, spædbørnsdødelighed og moderniseringsprocesser i Danmark 1800 til 1920*. Copenhagen: Gyldendal.

- Ludvigsen, Louise, Barbara Revuelta-Eugercios, and Anne Løkke. 2023. "Cause-Specific Infant Mortality in Copenhagen 1861–1911 Explored Using Individual-Level Data." *Historical Life Course Studies* 13 (January 17, 2023): 9–43. ISSN: 2352-6343, accessed October 30, 2024. <https://doi.org/10.51964/hlcs12032>. <https://hlcs.nl/article/view/12032>.
- Mandemakers, Kees, Gerrit Bloothoof, Fons Laan, Joe Raad, Rick J. Mourits, and Richard L. Zijdemann. 2023. "LINKS. A System for Historical Family Reconstruction in the Netherlands." *Historical Life Course Studies* 13 (June 1, 2023): 148–185. ISSN: 2352-6343, accessed September 6, 2024. <https://doi.org/10.51964/hlcs14685>. <https://hlcs.nl/article/view/14685>.
- Marker, Hans Jørgen. 2015. *Danmarks befolkning 1801, analyse på grundlag af folketællingen som mikrodata*. Odense: Syddansk Universitetsforlag.
- Ørberg, Paul G. 1991. *Hvad præsten skrev - i kirkebogen : kirkebøger og deres brug*. Arkivernes informationsserie. Copenhagen: Rigsarkivet.
- Park, Narae. 2022. "Record linkage of Norwegian historical census data using machine learning." Master Thesis, The Arctic University of Norway, August 2, 2022. <https://munin.uit.no/handle/10037/28399>.
- Revuelta-Eugercios, Barbara, Helene Castenbrandt, and Anne Løkke. 2022. "Older rationales and other challenges in handling causes of death in historical individual-level databases: the case of Copenhagen, 1880–1881." *Social History of Medicine* 35, no. 4 (November 1, 2022): 1116–1139. ISSN: 0951-631X, accessed May 20, 2024. <https://doi.org/10.1093/shm/hkab037>. <https://doi.org/10.1093/shm/hkab037>.
- Revuelta-Eugercios, Barbara, Olivia Robinson, Nicolai Mathiesen, Asbjørn Thomsen, Lise Sunde, and Anne Løkke. n.d. "Developing a method to create a benchmark dataset of linked historical individual data. Denmark, 1845–1901."
- Robinson, Olivia, Asbjørn Romvig Thomsen, Nicolai Rask Mathiesen, and Barbara Revuelta-Eugercios. 2023. "Transforming archival records into historical big data: Visualising human and computer processes in the Link-Lives project." In *The Nordic Model of Digital Archiving*. Routledge. ISBN: 9781003325406.
- Rosen, Wilhelm von, ed. 1983. *Rigsarkivet og hjælpemidlerne til dets benyttelse. 1,2: Handelskompagnier og guvernementer i de tropiske kolonier. Kgl. handel på Nordatlanten. Diverse erhvervsvirksomheder, Postvaesenet. Udskrivningsvaesenet. Landetaten. Sletaten. Regnskaber. Sønderjyske fyrstearkiver. Aeldre lokalarkiver ... / red. af Wilhelm von Rosen*. ISBN: 9788774970941.
- , ed. 1991. *Rigsarkivet og hjælpemidlerne til dets benyttelse. 2 Bd. 4: 1848 - 1990 Byggeri og boligforhold, fredning, naturforvaltning, fysisk planlægning, forurening, miljø, sundhedsvæsen, finansforvaltning, forsvarets civile myndigheder, kartlægning, særligt stillede landsdele, arkiver af fremmed proveniens, samling af kort og tegninger*. Vol. 4. København: Rigsarkivet [u.a.] ISBN: 9788774971290.

- Ruggles, Steven. 2002. "Linking historical censuses: a new approach." *History & Computing* 14 (1): 213–224. ISSN: 09570144. <http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=22850908&site=ehost-live>.
- Ruggles, Steven, and Diana L. Magnuson. 2020. "Census Technology, Politics, and Institutional Change, 1790–2020." *Journal of American History* 107, no. 1 (June 1, 2020): 19–51. ISSN: 0021-8723, accessed July 7, 2020. <https://doi.org/10.1093/jahist/jaaa007>. <https://academic.oup.com/jah/article/107/1/19/5862178>.
- Sköld, Olle. 2025. "The Concept of Paradata." In *Paradata: Documenting Data Creation, Curation and Use*, edited by Isto Huvila, Lisa Andersson, Olle Sköld, Ying-Hsang Liu, and Zanna Friberg, 11–39. Cambridge: Cambridge University Press. ISBN: 9781009366618, accessed December 11, 2025. <https://doi.org/10.1017/9781009366564.003>. <https://www.cambridge.org/core/books/paradata/concept-of-paradata/FE2EEBCEAF40CE622D7A85E1CC93535E>.
- Thomsen, Asbjørn Romvig. 2010. "Lykkens smedje?, social mobilitet og social stabilitet over fem generationer i tre jyske landsogne 1750-1850." PhD diss., University of Copenhagen.
- Thorvaldsen, Gunnar, Trygve Andersen, and Hilde L. Sommerseth. 2015. "Record Linkage in the Historical Population Register for Norway." In *Population Reconstruction*, edited by Gerrit Bloothoof, Peter Christen, Kees Mandemakers, and Marijn Schraagen, 155–172. (online): Springer.
- Vick, Rebecca, and Lap Huynh. 2011. "The Effects of Standardizing Names for Record Linkage: Evidence from the United States and Norway." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 44, no. 1 (January 31, 2011): 15–24. ISSN: 0161-5440. <https://doi.org/10.1080/01615440.2010.514849>. <http://dx.doi.org/10.1080/01615440.2010.514849>.
- Wisselgren, Maria J., Sören Edvinsson, Mats Berggren, and Maria Larsson. 2014. "Testing Methods of Record Linkage on Swedish Censuses." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 47, no. 3 (July 3, 2014): 138–151. ISSN: 0161-5440. <https://doi.org/10.1080/01615440.2014.913967>. <http://dx.doi.org/10.1080/01615440.2014.913967>.