link-
LIVES

Link-Lives Guide

version 1

June 2022

Olivia Robinson
Nicolai Rask Mathiesen
Asbjørn Romvig Thomsen
Barbara Revuelta-Eugercios

UNIVERSITY OF COPENHAGEN
Rigsarkivet
ancestry
CARLSBERG FOUNDATION
Innovation Fund Denmark

What is this guide?

This is a brief overview of how the Link-Lives project has prepared transcribed historical records and transformed them into linked research datasets. Further releases of the data and documentation will follow over the course of the project which runs until 2024.

Who is this guide for?

Researchers who wish to understand the contents of the dataset release v.1.2.1.

How to cite

Guide: Robinson, O., Mathiesen, N., Thomsen, A., Revuelta-Eugercios, B. (2022) *Link-Lives Guide v.1*, Danish National Archives/University of Copenhagen, Denmark

Data: Mathiesen, N., Robinson, O., Thomsen, A., Revuelta-Eugercios, B. (2022) *Link-Lives Data v.1.2.1*, Danish National Archives/University of Copenhagen, Denmark

Note that both resources can be downloaded at the National Archive's webpage and will be given a URL there which should be included in the citation.

Contact

While we welcome contact via the Link-Lives project website and general email address linklives@hum.ku.dk, please be aware that our resources do not stretch to technical or data support for users of Link-Lives datasets at this time.

Contents

1 Project & data description

1.1 What is Link-Lives?

Link-Lives is a cross-disciplinary research project that takes information relating to any given person, drawn from diverse archival sources, to build 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 multigenerational big linked data. The result will be a research infrastructure at the Danish National Archives, created in cooperation with the Copenhagen City Archives (Københavns Stadsarkiv) and the University of Copenhagen. // // Link-Lives is a five-year research project funded by Innovation Fund Denmark and the Carlsberg Foundation and runs until 2024. More information about the project scope, partners and participants can be found at link-lives.dk. A searchable, but not downloadable, version of the Link-Lives datasets can be accessed at link-lives.dk/soeg/ (in Danish only).

1.2 Overview of dataset contents

The Link-Lives datasets are a work in progress until 2024. The downloadable zip files contain Link-Lives datasets, licence files and a guide to the contents.

Link-Lives supplies two different types of data packages: a source type and a link type. The source type contains both the transcribed data in almost exactly the form it arrived into Link-Lives and the standardized Link-Lives version. There is one source type package for every source, i.e. 10 censuses, 6 parish records and 1 Copenhagen burial register package.

The link type package contains a mix of links (between pairs of sources), full life-courses (collections of links) and other supporting data, e.g. name and place standardization lists and tables with source IDs. This first data release contains life-courses linked across censuses (1787 to 1901), parish records (1813-1917) and Copenhagen burial registers (1861-1912).

2 Historical sources

The downloadable datasets are based on transcriptions of original, non-digital sources: national and city-wide population censuses (*folketællingerne*), parish records (*kirkebøgerne*) and the Copenhagen burial registers (*begravelsesprotokollerne*).

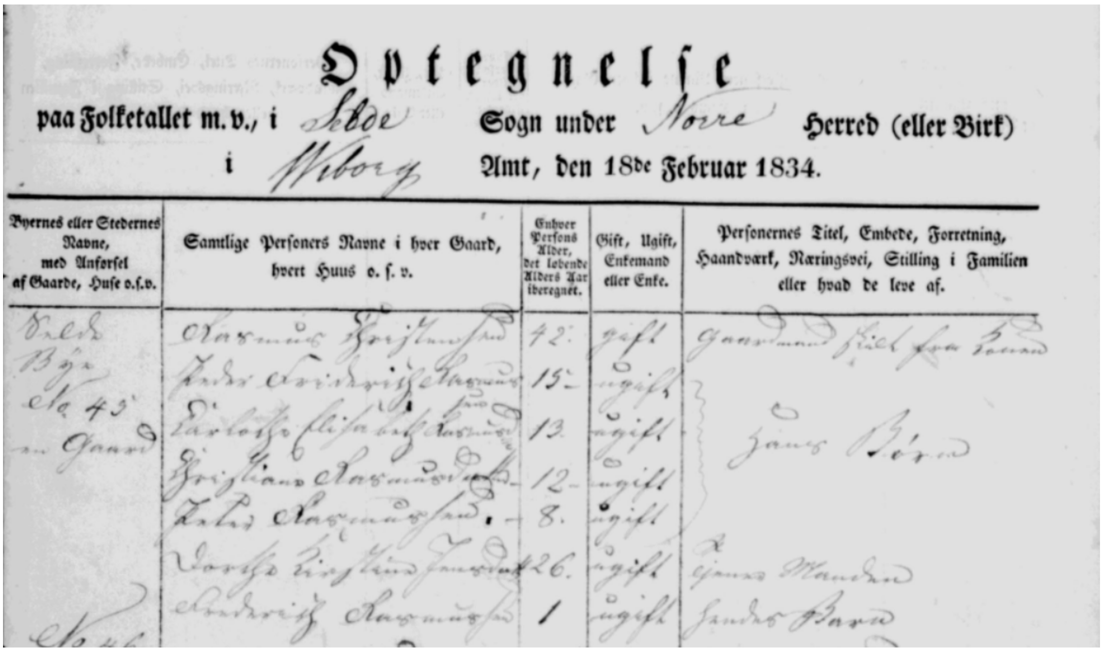
2.1 Censuses

A census is a snapshot of a population taken on a specific day. The nominal census lists contain basic demographic information about every individual in a household, such as names, ages, and occupations.

2.1.1 Census years

In Denmark, the first census was carried out in 1769, but apart from a few surviving nominal lists only the aggregated numbers from that census still exist. The first preserved nominal census is the 1787 census, followed by censuses in the years 1801, 1834, 1840, 1845, 1850, 1855, 1860, 1870, 1880, 1890, 1901, 1906, 1911, 1916, 1921, 1925, 1930, 1940, 1950, 1960 and 1970. Census-taking ceased once the Danish digital CPR-system (Central Person Registry) was established in 1968. City-wide censuses of the Copenhagen population were made in 1885 and 1895.

The Link-Lives datasets include the censuses of 1787, 1801, 1834, 1840, 1845, 1850, 1860, 1880, 1885 (Copenhagen only) and 1901. These censuses are the only ones that have been fully transcribed. More censuses will be added to the datasets as and when further censuses are transcribed.



Optegnelſe

paa Folketallet m. v., i Selde Sogn under Nørre Herred (eller Birk)
i Wiborg Amt, den 18de Februar 1834.

Byernes eller Stedernes Navne, med Anførſel af Gaarde, Huſe o.ſ.v.	Samtlige Perſoners Navne i hver Gaard, hvert Huus o. ſ. v.	Enhver Perſons Alder, det løbende Alders Aar iberegnet.	Gift, Ugift, Enkemand eller Enke.	Perſonernes Titel, Embede, Forretning, Haandværk, Næringsvei, Stilling i Familien eller hvad de leve af.
Selde By No 43 en Gaard	Rasmus Christensen	42.	gift	Gaardmand skilt fra Konen
	Peder Friderich Rasmussen	15.	ugift	Hans Børn
	Carlotte Elisabeth Rasmussen	13.	ugift	
	Christian Rasmussen	12.	ugift	
	Peder Rasmussen	8.	ugift	
	Dorthe Kirstine Jensdatter	26.	ugift	[illegible]
No 46	Andreas Rasmussen	1	ugift	Hendes Barn

Figure 1: An example of a household from the census of 1834: the household resides on a farm in Selde parish in Jutland, and it consists of a divorced man (42 years), his four children (8-15 years), a servant girl (26 years) and her son (1 year).

2.1.2 Census geography

The above mentioned censuses generally cover the population in Jylland, Fyn, Sjælland and adjacent islands. See a full breakdown of the geographic coverage of the transcribed datasets.

2.1.3 Census information

Census information gathered differs from year to year. Common for all censuses is name, age, civil status, relationship to other householders and occupation. From 1845 onwards, the parish of birth was recorded followed by more variables as time went on. The transcribed datasets used by Link-Lives contain all variables.

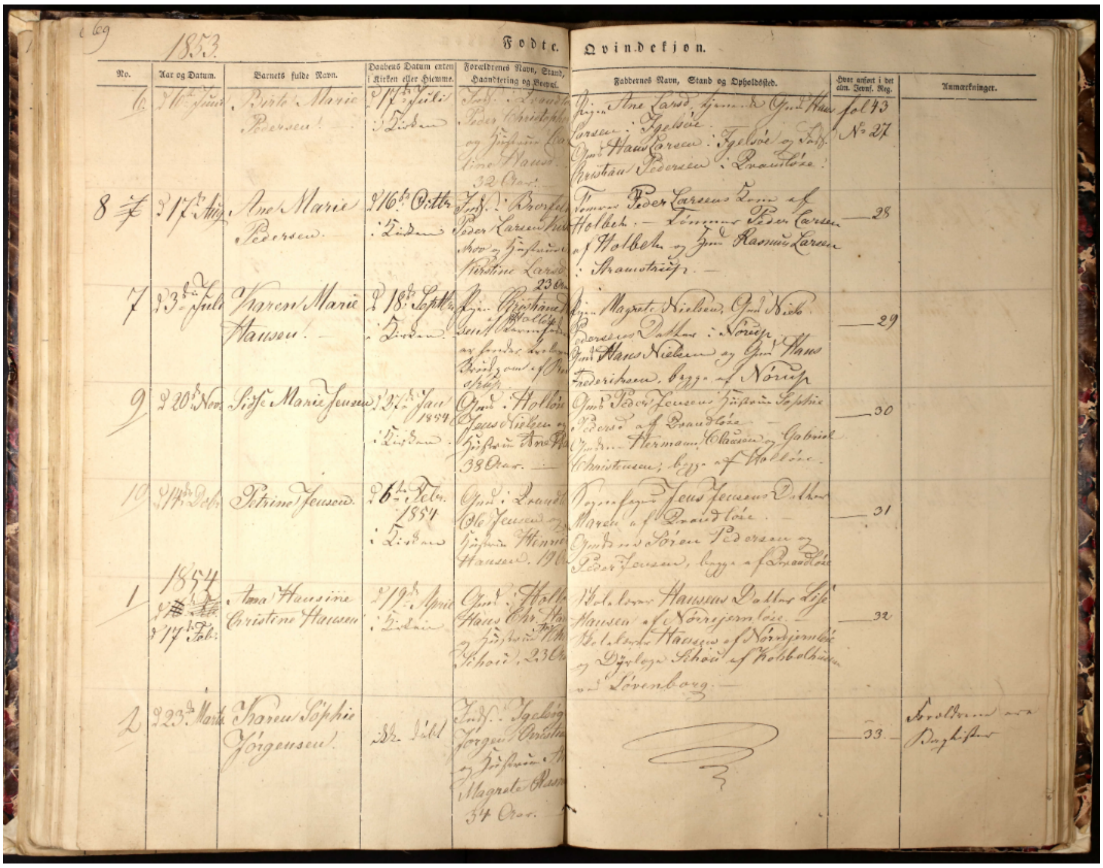
2.2 Parish records

Parish records are arranged chronologically, recording vital events and administered by the Danish church. The main events registered are baptisms, confirmations, marriages and burials. For a limited period, arrivals and departures of residents in and out of the parish were also registered.

2.2.1 Parish record years

From 1645, it was a task of the Established Church (Lutheran) parish priest to keep the parish records both for statistical and legal identification purposes. A major improvement occurred in 1813, when the local priests and parish clerks were instructed to keep two identical versions of the registers to safeguard against damage or destruction. From c.1813 onwards, therefore, at least one copy of parish records survive from each Danish parish. Arrivals and departures of certain parts of the population were to be recorded in the period 1813-1875, but in reality these lists are of varying quality and coverage.

The Link-Lives datasets include the parish registers from the period 1813-1917. The original records can be viewed on Ancestry's site (follow the instructions in Danish).



169

1853

Fødte. Qvindekjøn.

No.	Aar og Datum.	Barnets fulde Navn.	Daabens Datum enten i Kirken eller Hiemme.	Forældrenes Navn, Stand, Haandtering og Boepæl.	Faddernes Navn, Stand og Opholdssted.	Hvor anført i det alm. Jevnf. Reg.	Anmærkninger.
6.	d. 16de Juni	Birte Marie Pedersen.	d. 17de Juli : Kirken	Huusm. : Kvanløse Peder Christophersen og Hustru Caroline Hansd. 32 Aar.	Pige Ane Larsd. tjenende Gmd Hans Larsen : Igelsøe. Gmd Hans Larsen : Igelsøe og Ungk. Kristian Pedersen : Kvanløse.	fol 43 No 27	
8 ~~7~~	d. 17de Aug	Ane Marie Pedersen.	d. 16de Octbr : Kirken	Huusm. : Brorfelde Peder Larsen Kirkeskov og Hustru Kirstine Larsd. 23 Aar	Tømmer Peder Larsens Kone af Holbek. — Tømmer Peder Larsen af Holbek og Gmd Rasmus Larsen : Stramstrup. —	28	
7	d. 3die Juli	Karen Marie Hansen.	d. 18de Septbr : Kirken	Pige Christiane [illegible] [illegible] Holløse [illegible] Brudgom af [illegible]	Pige Magrete Nielsen, Gmd Niels Pedersens Datter : Nørup. Gmd Hans Nielsen og Gmd Hans Frederiksen, begge af Nørup	29	
9	d. 20de Nov.	Sidse Marie Jensen	d. 27de Jan 1854 : Kirken	Gmd : Holløse Jens Nielsen og Hustru Ane [illegible] 38 Aar.	Gmd Peder Jensens Hustru Sophie Pedersd. af Kvanløse — Gmd.. Hermann Clausen og Gabriel Christensen; begge af Holløse.	30	
10	d. 14de Decbr	Petrine Jensen.	d. 6te Febr. 1854 : Kirken	Gmd : Kvanløse Ole Jensen og Hustru Henriette Hansen. 19 Aar	Pigen Jens Jensens Datter Maren af Kvanløse. — Gmdrne Søren Pedersen og Peder Jensen, begge af Kvanløse	31	
1	1854 ~~d. [illegible]~~ d. 17de Febr.	Ane Hansine Christine Hansen	d. 19de April : Kirken	Gmd : Holløse Hans Chr. Hansen og Hustru [illegible] Schou. 23 Aar	[illegible] Hansens Datter Lise Hansen af Nørrejernløse. [illegible] Hansen af Nørrejernløse og [illegible] Schou af Kobbelhuus ved Løvenborg. —	32	
2	d. 23de Marts	Karen Sophie Jørgensen.	Ikke døbt	Indst. : Igelsøe Jørgen Christensen og Hustru Ane Magrete Rasmusd. 34 Aar.		33.	Forældrene ere Baptister

Figure 2: A page from a parish register containing baptised girls from Kvanløse parish 1853-1854. The columns contain information on the children's birth dates, names, baptism dates, parents' names and residence, and godparents' names and residence.

2.2.2 Parish record geography

The Danish parish records cover the entire contemporary territory of Denmark, including the southern part of Jutland, even in the years 1864-1920 under its German rule. The Link-Lives datasets include these. Although parish records were also made beyond these regions, they are not included in the Link-Lives datasets at present.

2.2.3 Parish records and religious communities

The vast majority of the Danish population in the period in question belonged to the Danish Lutheran State church. This part of the population is registered in the State Church parish records, and they are all included in the Link-Lives datasets.

A small part of the population (1850: 0.5%, 1901: 1.4%) belonged to other religious communities, such as Jewish, Roman Catholic or Methodist. These officially recognized religious communities were entitled to keep their own records. Some are included in the Link-Lives datasets but we do not know the exact proportion involved.

2.2.4 Parish record information

Baptism records contain dates of birth and baptism, the name and sex of the child, and names, residences and occupations of the parents and the godparents. From around the mid-19th century, the mother's age is registered and, later, even the birth and marriage dates and places of the parents. Confirmation records contain confirmation dates, name and age of the confirmand, names, civil status, residence and occupations of the parents, and in the earlier part of the 19th century also the priest's evaluation of the confirmand's behaviour and religious knowledge. Later, the date and place of birth the confirmand are both recorded. Marriage records contain the marriage date, the names, ages, civil status, occupations and residences of the bride and groom, and the name, occupation and residence of the best man. Later, the date and place of birth for the bride and groom, along with the names and occupations of their parents, were recorded.

Departures and arrivals generally only record the date of the departure/arrival, the name and age of the migrant, and the name of the parish to which they are moving to/from. The migrant's occupation is often mentioned. A general rule is that the information is more dense, the later in the period. The main spatial differences are that (a) the parish records from the southern part of Jutland tend to contain more information than in the rest of the country, and (b) the parish records from urban areas contain sparser details on individuals than in the rural areas. Unfortunately, far from all the information held in these records have been transcribed.

2.3 Copenhagen burial registers

The Copenhagen burial registers are nominal lists of burials in the Copenhagen cemeteries. The scope of the registers was to record the payment for each burial arrangement– i.e. they were never designed for population registry purposes. Nevertheless, the information they hold about buried individuals is rich and very close to that of the parish record burials (see above). The crowd-sourced transcriptions with links to the original handwritten records are held at the Copenhagen City Archive and can be accessed via the Københavns Stadsarkiv website.

2.3.1 Burial register years

The burial registers began in 1861 and are still in use. Prior to 1887 they did not cover Jewish, Roman Catholic or military cemeteries. The Link-Lives dataset covers the years 1861-1912.

2.3.2 Burial register geography

The burial registers were meant to record those buried in Copenhagen cemeteries. However, we have discovered that some people (often Copenhagen residents) appear in the registers despite being buried outside the city.

2.3.3 Burial register information

The burial registers contain the name and age of the deceased, death date, cause of death, burial date and place, and residence address at death. The records also contain information about the payment of the burial arrangements and the occupation of the deceased (or their spouse/parent). From 1913, the birth date and place of the deceased is also included. Not all variables are included in the Link-Lives dataset.

3 Link-Lives data preparation

3.1 Transcription & enrichment

Our datasets have been digitised for us (by creating machine-readable versions of non-digital records) and we then further prepare these through standardisation (improving interpretability by losing some variability) and enhancement (the creation of new variables). The transcribed/digitised datasets have therefore each been further 'treated' for Link-Lives purposes.

3.1.1 Nationwide population censuses

The nationwide censuses (1787-1901) were transcribed by volunteers at the National Archives Denmark and are held online in machine-readable datasets as part of the DDD (Danish Demographic Database).

3.1.2 Nationwide parish records

The parish records include baptisms, confirmation, marriages, burials, arrivals and departures. They were transcribed by a commercial company in Asia whose transcribers were non-native speakers or readers of Danish. The Link-Lives treated parish record datasets are not publicly available yet.

3.1.3 Copenhagen burial registers

The Copenhagen burial protocols (1861-1942) were transcribed via a crowdsourcing project at the Copenhagen City Archives and are freely available to search at www.kbharkiv.dk. The Link-Lives treated Copenhagen burial protocols are not publicly available yet.

3.2 Standardization

A number of variables have been standardised to make them comparable. Names of persons and places are often spelled in different ways in historical sources, and therefore it is necessary to standardize them in order for them to be understood as two spelling variations of the same name during the linking process. Some, but not all, variables have been standardized. This has been relatively easy with sex, civil status, and age as these variables are standardized into either rather few or rather objectively definable standard spellings. A more complex approach was followed to account for name and place variations.

3.2.1 Person Names

All given and family names were extracted from the transcribed census datasets 1787-1901 and split out into individual strings (e.g. “Hans Christian Andersen” became “Hans”, “Christian”, “Andersen”). This resulted in 350,000 unique names, the majority of which appeared only once or very few times. For efficiency reasons, only the most frequently appearing names were then manually standardized: 6,233 name strings, representing c.95% of all occurrences in the censuses, were coded by two historians with expert domain knowledge of the original sources and linking. The resulting synonym catalogue was then confirmed by a researcher in onomastics (the study of names).

3.2.2 Birth places

Places of birth were recorded in the original sources in a free-text field called ‘place of birth’. All unique, original place name strings (over 600,000) were extracted from this field in the transcribed census datasets 1845-1901 and split out into individual words (eg. “Ølstrup Sogn, Ringkjøbing Amt” became “Ølstrup”, “Sogn”, Ringkjøbing”, “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 standardized: 5,000 name strings, representing c.97% of all occurrences in the censuses, were coded by a historian with expert domain knowledge of the original sources and linking. Key words such as “Sogn” or “S” (*parish*) and “Amt” or “A” (*county*) appearing in the original strings were used for classifying the unique place words as names of parishes, counties etc. Values such as “ditto” were replaced with the value of the preceding record, and values stating birth place as ”her i sognet” (*in this parish*) were replaced with the value in the variable “event parish”.

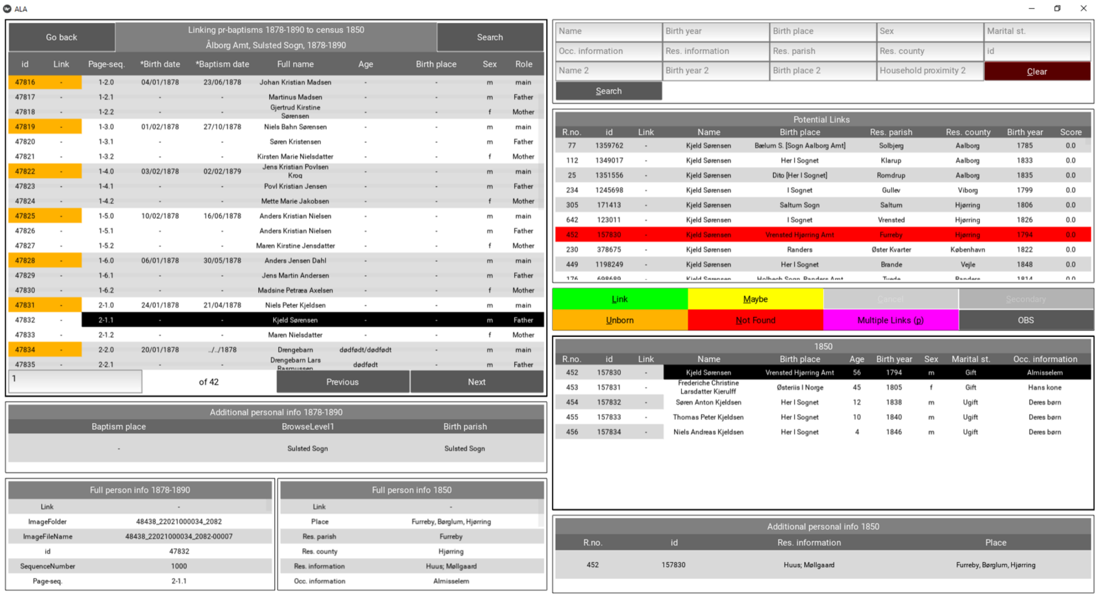
4 Linking methods

We include two methods of linking: manual domain expert links and automatic rule-based links. We link backwards in time, i.e. from the newer source to the older source, except for baptisms and marriages which are linked forward in time. The domain expert links (see section 4.1) are made by humans (domain experts) and only cover a small subset of the sources. To ensure representativity, we have selected linking units that are geographically scattered and socially diverse. The domain expert links have been created to function as high quality benchmark data to test the automatic algorithms against. We also plan to use the manual links as training data for machine-learning based linking algorithms. Although they have been created for technical purposes, we believe this limited set of links is of very high quality and represents a valuable research resource on its own. The rule-based links (see section 4.2) are made automatically by implementing a set of rules for a computer to follow. This link set has nationwide coverage.

4.1 Domain expert links

4.1.1 Linking interface

Each dataset is broken down into a linking unit, usually a parish or part of a parish. This is loaded into a purpose-built offline interface (ALA, see figure 3) in which linkers are presented with transcribed data from two sources and a subset of potential link candidates generated by a rule-based algorithm, using some relatively simple rules and standardizations. They can further refine their searches manually outside of the potential cases proposed, then make a linking decision based on the data presented to them on the screen.



Linking pr-baptisms 1878-1890 to census 1850
Ålborg Amt, Sulsted Sogn, 1878-1890

id	Link	Page-seq.	*Birth date	*Baptism date	Full name	Age	Birth place	Sex	Role
47816	-	1-2.0	04/01/1878	23/06/1878	Johan Kristian Madsen	-	-	m	main
47817	-	1-2.1	-	-	Martinus Madsen	-	-	m	Father
47818	-	1-2.2	-	-	Gjertrud Kirstine Sørensen	-	-	f	Mother
47819	-	1-3.0	01/02/1878	27/10/1878	Niels Bahn Sørensen	-	-	m	main
47820	-	1-3.1	-	-	Søren Kristensen	-	-	m	Father
47821	-	1-3.2	-	-	Kirsten Marie Nielsdatter	-	-	f	Mother
47822	-	1-4.0	03/02/1878	02/02/1879	Jens Kristian Povlsen Krog	-	-	m	main
47823	-	1-4.1	-	-	Povl Kristian Jensen	-	-	m	Father
47824	-	1-4.2	-	-	Mette Marie Jakobsen	-	-	f	Mother
47825	-	1-5.0	10/02/1878	16/06/1878	Anders Kristian Nielsen	-	-	m	main
47826	-	1-5.1	-	-	Anders Kristian Nielsen	-	-	m	Father
47827	-	1-5.2	-	-	Maren Kirstine Jensdatter	-	-	f	Mother
47828	-	1-6.0	06/01/1878	30/05/1878	Anders Jensen Dahl	-	-	m	main
47829	-	1-6.1	-	-	Jens Martin Andersen	-	-	m	Father
47830	-	1-6.2	-	-	Madsine Petræa Axelsen	-	-	f	Mother
47831	-	2-1.0	24/01/1878	21/04/1878	Niels Peter Kjeldsen	-	-	m	main
47832	-	2-1.1	-	-	Kjeld Sørensen	-	-	m	Father
47833	-	2-1.2	-	-	Maren Nielsdatter	-	-	f	Mother
47834	-	2-2.0	20/01/1878	../../1878	Drengebarn	dødfødt/dødfødt	-	m	main
47835	-	2-2.1	-	-	Drengebarn Lars Rasmussen	dødfødt	-	m	Father

Additional personal info 1878-1890

Baptism place	BrowseLevel1	Birth parish
-	Sulsted Sogn	Sulsted Sogn

Full person info 1878-1890

Link	-
ImageFolder	48438_22021000034_2082
ImageFileName	48438_22021000034_2082-00007
id	47832
SequenceNumber	1000
Page-seq.	2-1.1

Full person info 1850

Link	-
Place	Furreby, Børglum, Hjørring
Res. parish	Furreby
Res. county	Hjørring
Res. information	Huus; Møllgaard
Occ. information	Almisselem

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]	Solbjerg	Aalborg	1785	0.0
112	1349017	-	Kjeld Sørensen	Her I Sognet	Klarup	Aalborg	1833	0.0
25	1351556	-	Kjeld Sørensen	Dito [Her I Sognet]	Romdrup	Aalborg	1835	0.0
234	1245698	-	Kjeld Sørensen	I Sognet	Gullev	Viborg	1799	0.0
305	171413	-	Kjeld Sørensen	Saltum Sogn	Saltum	Hjørring	1806	0.0
642	123011	-	Kjeld Sørensen	I Sognet	Vrensted	Hjørring	1826	0.0
452	157830	-	Kjeld Sørensen	Vrensted Hjørring Amt	Furreby	Hjørring	1794	0.0
230	378675	-	Kjeld Sørensen	Randers	Øster Kvarter	København	1822	0.0
449	1198249	-	Kjeld Sørensen	Her I Sognet	Brande	Vejle	1848	0.0
176	698689	-	Kjeld Sørensen	Holbech Sogn, Randers Amt	Tvede	Randers	1814	0.0

1850

R.no.	id	Link	Name	Birth place	Age	Birth year	Sex	Marital st.	Occ. information
452	157830	-	Kjeld Sørensen	Vrensted Hjørring Amt	56	1794	m	Gift	Almisselem
453	157831	-	Frederiche Christine Larsdatter Kierulff	Østeriis I Norge	45	1805	f	Gift	Hans kone
454	157832	-	Søren Anton Kjeldsen	Her I Sognet	12	1838	m	Ugift	Deres børn
455	157833	-	Thomas Peter Kjeldsen	Her I Sognet	10	1840	m	Ugift	Deres børn
456	157834	-	Niels Andreas Kjeldsen	Her I Sognet	4	1846	m	Ugift	Deres børn

Additional personal info 1850

R.no.	id	Res. information	Place
452	157830	Huus; Møllgaard	Furreby, Børglum, Hjørring

Figure 3: The linking interface (ALA) with options to browse data, search the sources and make link-decisions.

4.1.2 Link-Lives linkers

Each linker is trained in the Link-Lives approach through a training process called Linking School. After linking 3 previously-linked units to a satisfactory standard, they join the rest of the production team. We support linkers through monthly Link-In workshops (virtual or in-person sessions to share knowledge and solve linking challenges) and an online chat facility for remote support. In total, over 30 people have been trained to link consistently using the ALA software and a core team of 8-10 linkers regularly link parish records and census records to census records, to build high-quality training data. Linker ages range from 20s to 60s, they are both male and female, historians and non-historians, paid and unpaid (volunteers).

4.1.3 Key features of the Link-Lives linking approach

Linkers are guided by a number of linking principles, governed by one key rule: make a linking decision for every person in a given linking unit.

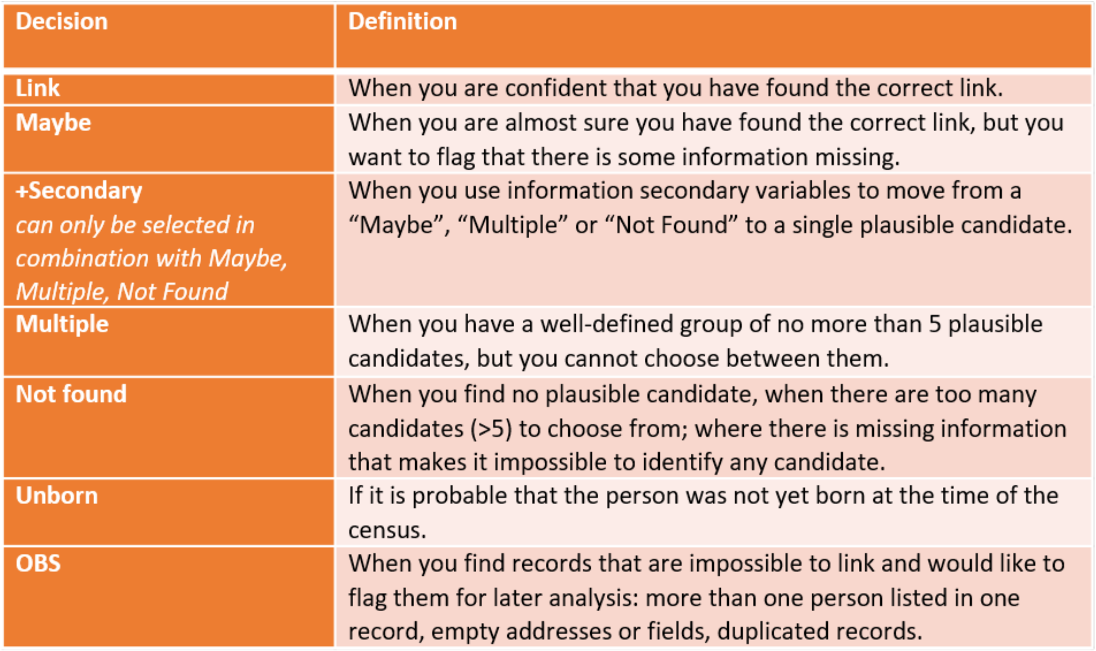
- Despite the support we provide to guide and align their linking decisions, no two linkers link in exactly the same way. We require each record to be linked by two linkers, who then get a chance to review any links where they did not agree.

- Linkers first assess primary variables (information thought to change little - or predictably - over time) to inform the likelihood of a link: name, birth year, birth place, gender, family/household context, civil status. Only after making a decision based on these do linkers use any other information (i.e. secondary variables). This reduces bias potential.

- To resolve any enduring conflicts (where linkers continue to disagree after reviewing their links), a third independent linker then resolves those cases where linkers continue to disagree (c.5-15% of cases). The file is then saved as a Consolidated Training Data (CTD) file.

4.1.4 Linking decisions

Positive links (plausible matches) attract a Link or a Maybe. Negative links (not enough information is available to make a plausible link) are marked Multiple or Not Found. OBS and Unborn indicate that a record is for some reason impossible to link. +Secondary indicates where non-primary information has been used to make the link more secure. The definition of each link decision is described in figure 4.



Decision	Definition
Link	When you are confident that you have found the correct link.
Maybe	When you are almost sure you have found the correct link, but you want to flag that there is some information missing.
+Secondary *can only be selected in combination with Maybe, Multiple, Not Found*	When you use information secondary variables to move from a “Maybe”, “Multiple” or “Not Found” to a single plausible candidate.
Multiple	When you have a well-defined group of no more than 5 plausible candidates, but you cannot choose between them.
Not found	When you find no plausible candidate, when there are too many candidates (>5) to choose from; where there is missing information that makes it impossible to identify any candidate.
Unborn	If it is probable that the person was not yet born at the time of the census.
OBS	When you find records that are impossible to link and would like to flag them for later analysis: more than one person listed in one record, empty addresses or fields, duplicated records.

Figure 4: Definition of link decisions.

These link decisions are made on each person appearance based on the flowchart guide seen in figure 5.

4.1.5 Link rates

The link rates for manual linking vary especially according to source type (see table). Sources featuring more of the primary variables are more readily linked (eg. late-1800s censuses). High contention rates (where linkers disagree more often) also result in lower link rates.

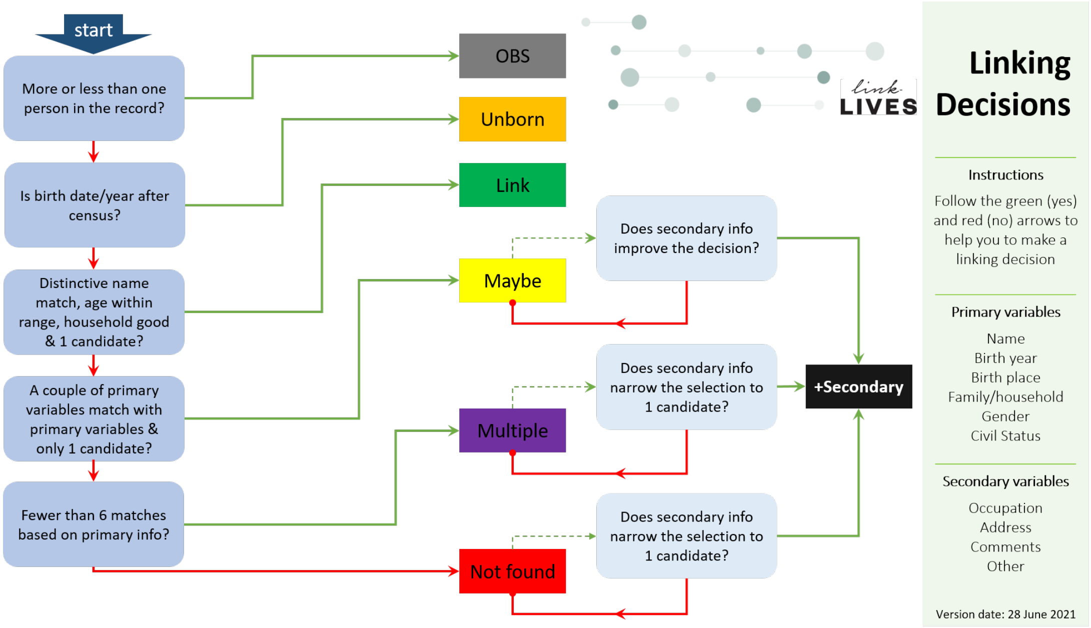

start
More or less than one person in the record?
Is birth date/year after census?
Distinctive name match, age within range, household good & 1 candidate?
A couple of primary variables match with primary variables & only 1 candidate?
Fewer than 6 matches based on primary info?
OBS
Unborn
Link
Maybe
Multiple
Not found
Does secondary info improve the decision?
Does secondary info narrow the selection to 1 candidate?
Does secondary info narrow the selection to 1 candidate?
+Secondary
link LIVES
Linking Decisions
Instructions
Follow the green (yes) and red (no) arrows to help you to make a linking decision
Primary variables
Name
Birth year
Birth place
Family/household
Gender
Civil Status
Secondary variables
Occupation
Address
Comments
Other
Version date: 28 June 2021

Figure 5: Flowchart guiding the link decision making.

4.2 Rule-based links

The rule-based (RB) links are based on a set of rules determining what constitutes a link. That means that every potential pair of records (link) have been assigned a score which determines the quality of the link. Following this, a threshold value has been set in an attempt to sort *correct* links from *incorrect* links. In this section we will 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.

4.2.1 The algorithm

The algorithm includes 5 major steps:

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

Step 4 and 5 are repeated until no further links are produced by repeating. After reaching convergence, the threshold parameters are adjusted to allow for more uncertain links to be created, thus repeating the process until all preset thresholds have been used. See figure 6 for visual representation of the algorithm.

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 *potential* links to consider, 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 ±2 years to pass the blocking stage. Due to the age criteria, anyone with no age (or birth year) is at present not considered for linking. This means that most often, only the *main* person(s) on the PRs can be linked at this time.

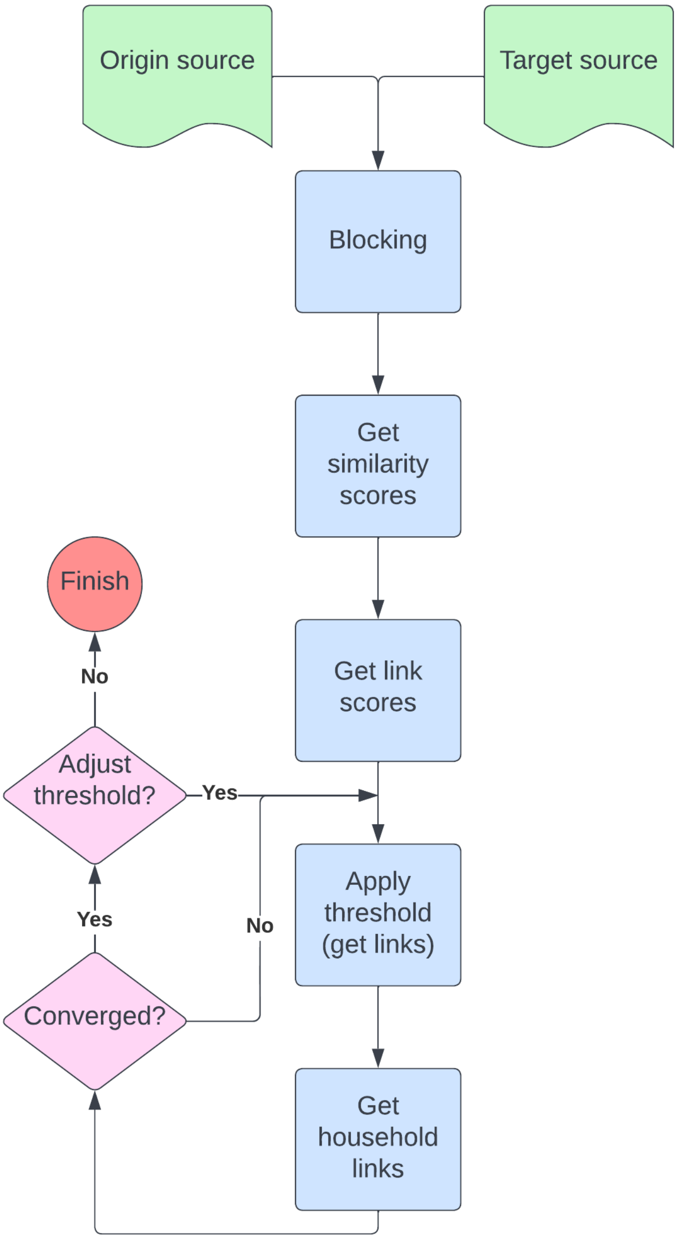


Figure 6: The rule-based algorithm.

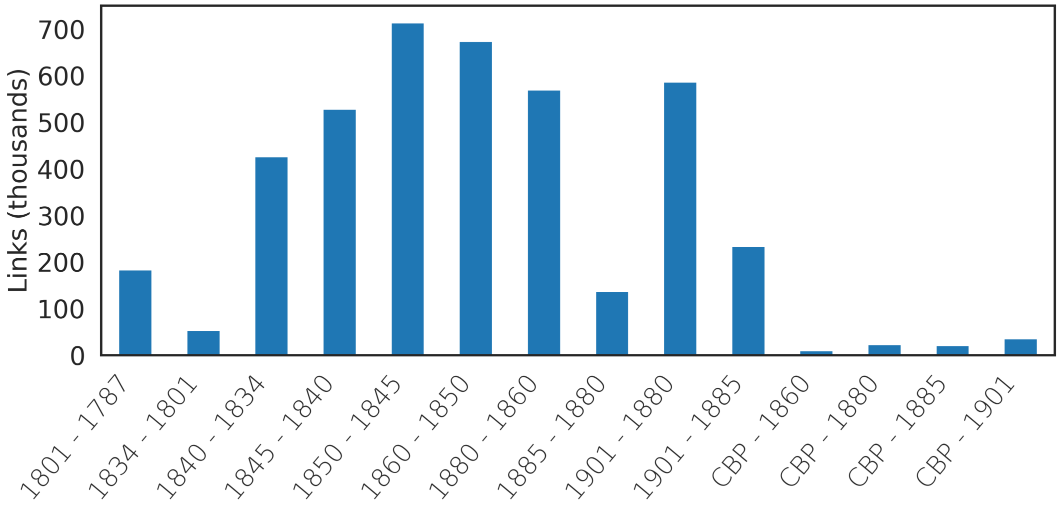


Figure 7: Total count of links from census to census and from Copenhagen burial register to census.

Similarity scores We use the Jaro-Winkler distance for comparing name and birth place strings. For names, the score is based on comparing the standardized and categorized first names, family names and patronyms. For the birth place, the score is based on the standardized and categorized birth place variables.

Link score The link score is simply the sum of similarity scores. Since we have used the Jaro-Winkler distance, a score of zero will equal a prefect match.

Applying thresholds (primary links) After calculating link scores for all potential links passing the blocking stage, it is then time to decide what is considered a (primary) link. 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, any matching married couples between the households are used to disambiguate the decision.

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.

Iteration and tolerance adjustment After having made household links, some of the potential links that were ambiguous before, now possibly standout 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 over again.

4.2.2 How many links?

There is currently a total of 5.5 million rule-based links in the database. The distribution can be seen on figure 7 and 8 and the link rates are shown in 9. The figures show that the most links are made in the period 1845-1901 (the censuses featuring individuals' birth places) while the early period 1787-1845 isn't as well represented.

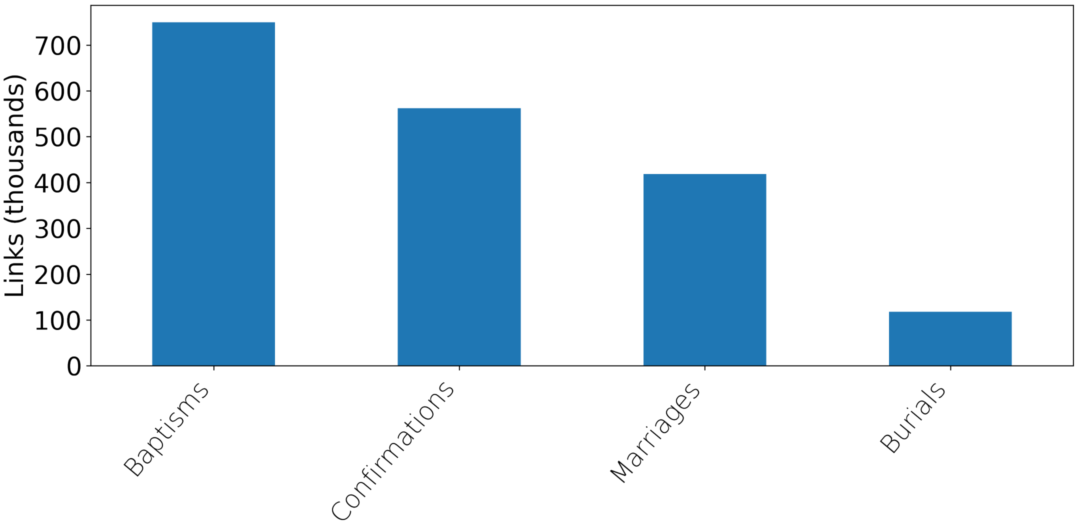


Figure 8: Total count of links from PR to censuses.

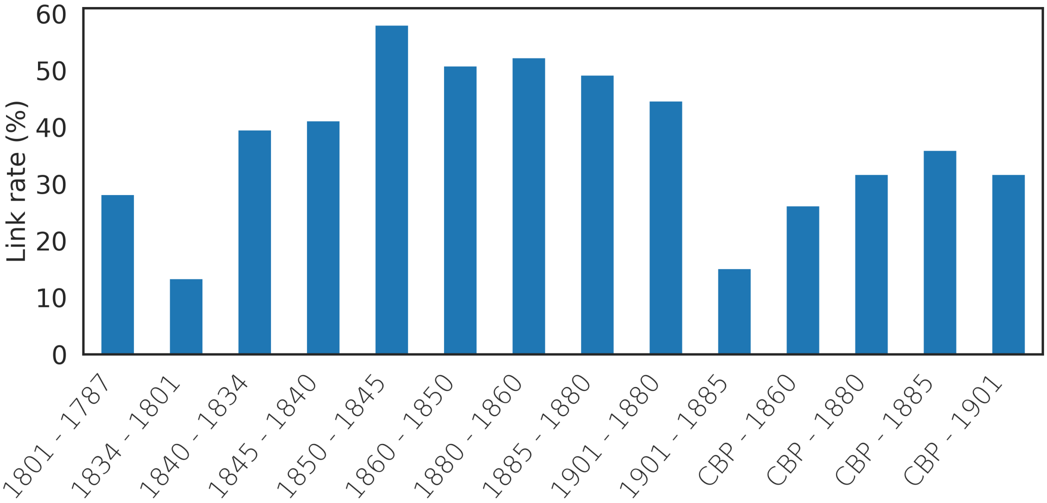


Figure 9: Link rates of links made from censuses and Copenhagen burial register.

4.2.3 What is the quality?

We can measure the quality of the RB links by comparing them with the domain-expert (DE) links. For now, most of our DE links are focused on the censuses and specifically on the period 1845-1901. 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%. In future versions of this document, we will extend this section with more in-depth analysis and detailed information on how to understand these numbers further.

5 Life-courses

After having linked all sources to the nearest census, we chain the links together into life-courses. A life-course can be the trivial case of a single link up to many links including records from the censuses, parish records and Copenhagen burial registers.

The overall strategy is to identify *end points*, i.e. records which are only connected in one direction, then build the life-course by connecting the links iteratively. 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. The algorithm will be described in greater detail in a later version of this document. Currently there are more than 3 million life-courses in the Link-Lives dataset with up to 12 person appearances.