Link-Lives, Historical Big Data: Reconstructing Millions of Life Courses from Archival Records Using Domain Experts and Machine Learning*

Bárbara A. Revuelta-Eugercios[1, 2] [0000-0002-2449-037X], Olivia Robinson[2] [0000-0003-3085-0025] and Anne Løkke[2]

[1] *Rigsarkivet*/National Archives of Denmark, (Jernbanegade 36A, Odense, 5000, Denmark)
[2] University of Copenhagen (SAXO Institute, Department of History, Karen Blixens Plads 4, Copenhagen 2300, Denmark)
bre@sa.dk

Abstract. The Danish archives comprise some of the world's most comprehensive source coverage but despite large-scale digitization and transcription projects by diverse actors, there are no shared standards or possibilities for data linkage. The Denmark-based Link-Lives research project (2019-2024) is tackling this disparity by linking individual-level Danish records in census and parish record sources from 1787-1968 to create a multigenerational database for research using a combination of domain expertise and machine learning techniques. In contrast to small-sample linking or fully automated processes, Link-Lives is creating its own manually-linked data to train machine learning as well as exploring the impacts of different approaches to linking. Due to personal data protection legislation and propriety agreements, the data cannot be fully open access, but data outputs will be made available to both researchers and the general public via a website. The project's interdisciplinary team is based at the Danish National Archives and the University of Copenhagen, in partnership with Copenhagen City Archives, and funded by Carlsberg and Innovation Fund Denmark.

Keywords: archival record, multigenerational data, big data, record linkage

1 Introduction

The Danish archives comprise some of the world's most comprehensive source coverage of the lives of individuals but despite large-scale digitization and transcription projects by diverse actors, there are no shared standards or ready-made possibilities for data linkage. Link-Lives is a cross-disciplinary research project that combines information relating to any given person drawn from diverse archival sources, to build life courses and family relations from 1787 to the present. We combine machine learning, historical research, bioinformatics, 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. It will expand the scope of registry-based research from decades to centuries, and opens up new avenues for intergenerational research in the health and social sciences. Denmark

will be first in the world to implement this level of digitalization in full functional scale as part of the service of its national archives. The results will be made available for two main user groups: to researchers across disciplines, from historians to health and social scientists, who will use the life-course and multigenerational data to explore new research areas; and to the public, via a website disseminating the part of the data that is unrestricted by legislation or propriety agreements. The links and life courses will be freely available for anyone to search. Funded by the Innovation Fund Denmark and the Carlsberg Foundation, the project started in 2019 and will be completed in 2024.

The paper presents the aims and scope of the project along with some preliminary results. The structure is as follows: Section 1 provides an overview of the Danish context and how new projects are taking fresh advantage of the scope of archival digitization. Section 2 discusses the methodological developments to deal with sources from widely different provenances. Specifically, it focuses on the standardization and the innovations that derive high quality linked data to train machine learning approaches. Section 3 discusses our decisions on disseminating the linked data in light of current constraints and opportunities. Section 4 provides a conclusion and further perspectives.

1.1 Why do we need historical big data?

The biological and social life of humans is influenced by multigenerational mechanisms. However, these mechanisms are poorly understood, not least due to the enormous workload required to establish life courses and family relations spanning 4-5 generations [1]. In fact, for the era before the introduction of computerized civil registration systems (in Denmark, the CPR in 1968), most of our knowledge is based on limited samples: either long family pedigrees reconstructed manually within small geographic areas or reconstructions of just a few generations for larger areas. Research using even these limited datasets has nevertheless revealed promising findings: four hundred years of parish records in Canada showed selective fertility advantages on the French Canadian frontier [2]; rich contextual data revealed a transgenerational response to paternal grandfathers' access to food in 19th century Northern Sweden [3]; and a correlation was identified between political affiliation in the 2010s elections and cross-ethnic exposure in the early twentieth century in some states in the US [4], to cite some examples. In fact, the promise of new discoveries has led to new and renewed investments in the construction of large-scale databases that are pioneering this type of research. National-level work is underway in the Netherlands [5], Scotland [6], Norway [7], Sweden [8], the US [9] and Canada [10], while others are harvesting somewhat unrepresentative population-scale family trees from genealogy sites [11] using similar research goals.

1.2 The Danish context

The preconditions for this type of research have existed in Denmark for the last couple of decades. Rich archival material documenting many dimensions of individual lives over a period of more than 200 years has been preserved, and dedicated projects to transcribe and/or digitize many of these have been underway for some time. However, there has been relatively little research using these sources, as most have had to rely on

using small manually-linked samples [12, 13]. This has changed due to the advances in record linkage technologies, increased access to computing power in both personal computers and high performance computing facilities, and a growing awareness of the opportunities offered by these large untapped collections of historical data for many fields of research.

Revuelta-Eugercios, one of the Link-Lives PIs, was granted a project in 2014 that became a methodological pilot for what would later became Link-Lives, with a focus on both infrastructure building and research. More recently, the research potential of the area of large-scale historical population databases has been acknowledged in Denmark with three further projects attracting funding. Two of them will use and connect directly to Link-Lives data (a PhD and a collaborative project); a third, the Multigenerational Registry (funded by Novo Nordisk Foundation in 2021), is independent from Link-Lives. The MGR will make even more data available for the 20th century by transcribing parish records for the period 1920-1968, using image recognition to establish parental information for all individuals born after 1920, before linking them to the CPR registry.[14] The synergies between these two project investments are clear. In a few years, it will be possible to create a Danish Historical Population Registry, featuring life courses and relationships, as well as social, economic, and medical conditions, from a wide variety of sources. It will provide multigenerational data with long coverage, from 1787 to 1968, and high density, covering many dimensions of individual persons. Projects such as these are timely and poised to harness the power of both the depth of source coverage and breadth of digitization, to unleash the full potential of their research use.

2 Source material and methodological developments

2.1 Obtaining data through a wide-partnerships: old and new crowdsourcing data and collaboration with Ancestry

In the past twenty years, research projects, crowdsourcing initiatives, genealogists' homepages, and private companies have transcribed parts of the Danish archival heritage for their own use. Each project has invented ways of transcribing, standardizing and storing data, and developed their own ontologies, terminologies and data formats. Millions of records have been digitized, though with no shared standards or possibilities of linkage. These 'information islands' have excellent data, but it is both difficult and expensive to navigate the many databases and websites in the search of it, and any reuse of the material for different purposes requires expensive re-processing. Among the most important actors are volunteer-driven digitation (scanning/photographing from analogue to facsimile) and datafication (transcription into machine-readable format) initiatives which, in cooperation with archives, have made millions of records containing individual information available. The National Archives holds the oldest crowdsourcing heritage data project in Denmark (founded in 1992) [15] but many other archives, such as the Copenhagen City Archives [16] and Aarhus City Archives, have

started similar projects in the last decade. Genealogical associations have also participated in recent years and private companies have begun to enter the field with large digitation ventures of their own, with varying types of collaboration and data sharing agreements at both the national and municipal archive level. Additionally, there are some key examples of historical sources transcribed for contemporary health research [17, 18].
The current Link-Lives project (2019-2024) is linking four types of sources from four different actors with overlapping and linkable chronologies: 1) more than 20 million records, including the full count of 9 nationwide census years plus the local Copenhagen census of 1885 from the National Archives, plus 1921 and 1940, which will be newly transcribed by commercial actors; 2) 300,000 burials for the period 1861-1920 from a crowdsourcing project at Copenhagen City Archives; 3) 22 million parish records indexed by the genealogy company Ancestry; and 4) the Danish Civil Registration System (CPR), established in 1968. Further collections can and will be added to extend the range, subject to additional funding, since the infrastructure is being built to absorb new datasets.

Given the different origins, aims, and context of each of the collections, a key feature of the process is standardization to ensure that spellings and conventions are consistent with one another. In a first phase, we have developed synonym catalogues for historical names, places, civil status, etc. which ensure the maximal proximity to the source while taking into account existing vocabularies. In a second phase, we are mapping variables to existing classification schemes (and international ones where possible) and authority lists (if they exist). To build these, we employ domain experts, who undertake the slow yet important process of compiling synonym catalogues in combination with computer-assisted processes. Our aim is to provide standardized and simpler access to the vast collections, while still respecting the nature of the records and the context of their creation, preservation, and survival, through systematic documentation of metadata. This will ensure users can clearly identify what version of a variable they are working with and the process it has undergone prior to its coded form.

As of August 2021, we have completed the standardization of several sets of categorizations – those for given names, surnames, geographic places and causes of death – and we are currently working on coding occupations. We extracted millions of given and family names from census data and coded the 6,233 most-often occurring, which represent *c*.95% of all names. As there is no authority list for names, we collaborated with a scholar of onomastics to create our own. We have also coded over 600 unique causes of death, which covers 7,6% of all the unique causes of death that occurred in Copenhagen in the years 1861-1911. Despite the small number, this has allowed us to code 90% of all the deaths in Copenhagen during the period. The classification we use is the ICD10-h (International Classification of Disease 10 adapted for Historical purposes), which, as a partner in the SHiP project, we are currently involved in developing [22], with contribution from history of medicine specialists. Places of birth appear in the datasets in free text fields, in which places are recorded with all sorts of granularities so there are more than 600.000 unique strings for places. Out of 13,261 place-name

components that appeared in the place of birth fields in the 1845-1901 censuses, we have standardized 4,999 uniquely-spelled place-name words. These represent *c*.98% of all place-name occurrences for this period. We have derived a reference list too, for words that relate to existing places, from the DigDag project [21], which contains the most updated survey on geographical boundaries in Denmark from 1500 to today. While the residence parishes have already been geocoded to DigDag, we aim to geocode the actual standardized places and will start coding occupations to HISCO [19] in the fall of 2021.

2.2 Automatic methods in record linkage: domain-expert grounded algorithms

Until recently, a huge amount of manual effort was required to establish these linked life courses, but in the past decade, researchers in different countries have pioneered methods for automatic linkage on national scales (Norway, Sweden, Scotland, Netherlands, Canada and in the US) [7, 23–27]. The core challenge in creating intergenerational data from historical sources is entity resolution: identifying the same person across multiple sources (in a context of low identifiability of persons and high variation in quality and availability of critical attributes). Many of the existing automatic linking projects have mostly used rule-based linkage methods [7, 28]. In recent years, some researchers have carried out promising experiments with supervised machine learning techniques, including support vector machine and random forest [24–26, 29]. Some authors have even argued for the development of truly "automatic" probabilistic methods with little input from the researcher, to create simple, transparent and replicable methods that are mostly context-neutral [30]. However, their focus on the ease of implementation comes at the expense of the contextual, archival, historical features of the data. These overly-automated processes omit a thorough discussion of "ground truth" or "training data" for either testing or training models, which is a deficit that also applies to much of the historical record linkage literature. "Ground truth" is the term used to refer to information or data known to be real or true, provided by direct observation and/or measurement, and acts as a benchmark for any replication attempt. Training data usually refers to the specific dataset that is used to train machine learning models, which in the best of cases is ground truth. The challenge in developing historical record linkage approaches is that ground truth data does not actually exist: there is no additional source to verify whether a link between any two historical records is right or wrong, so all links are in reality the result of an estimation. Each researcher thus has to find and/or create their own "training data" to test or train their models.

The three more common practices for linking have been: harvesting linked records from genealogy sources; using manually-linked data produced by a single historian; or creating a small number of sets as part of a semi-automatic algorithm. Whatever the method chosen, though, researchers do not often provide information on the process of data construction or its consequences. The Link-Lives approach differs in this. Not only do we create our own manually-linked data, but we explore the impacts of different approaches. We take into account the specificity of data sources and contexts as well

as designing a protocol to identify the variation in human decision-making. We believe that putting history and historians at the center of linking methodologies creates data of the highest quality which is a pre-requisite to any successful machine learning model. We ensure that our training data is created by a three-domain-expert approach. This gives us assurances that any algorithm we create is built on robust data that reflects in a transparent way our best estimate at ground truth. Additionally, rather than committing to one specific type of model (which will always have its own strengths and limitations and is suited to specific research questions) Link-Lives is committed to develop or implement multiple models, in order to provide a variety of options for researchers.

We have used Python and Kivik to develop a bespoke interface named ALA (Assisted Linking Application) which requires no installation or programming competences, so it can be widely distributed to non-experts in programming. 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 then make a linking decision based on the data presented to them on the screen and can further refine their searches manually outside of the potential cases proposed. In total, over 30 people have been trained to link consistently using the ALA software and a core team of nine linkers is currently using it to link parish records and census returns to census data, to build high-quality training data. As at August 2021, the team has created 31,880 training data records.

Although linkers are governed by a set of best practices and attend linker workshops to guide and align their linking decisions, no two linkers link exactly the same way. We require each record to be linked by two linkers, before a third resolves those cases where linkers do not agree (*c*.5-15% of cases). We have also documented that linking rates vary widely by factors unrelated to linker experience or abilities. For example, our domain-expert linked sets show that in some rural parishes we achieve 95% link rates compared to 30-40% in urban areas (often explained by higher population mobility). The preliminary results from our relatively simple rule-based algorithm are somewhat lower but on a par with what we see in the international literature. On average, our early linking of the 1845, 1850 and 1860 census years shows a 50% link rate, but we expect other sources and later censuses to vary in this. We are now working on updated rule-based models but also have reached a sufficient threshold of manually linked data to start testing machine learning approaches. We are experimenting with variational recurrent neural networks, support vector machines and random forest specifications. Given the high computing needs for comparing millions of records while handling personal data protected by GDPR (individuals born after 1901), we use a cloud cluster in the Computerome 2.0 Supercomputer at Danish Technical University where we have a reserved node with 40 cores. This allows us to run multiple variations of the algorithms and test the effects on the linked data. For example, in the latest configuration, each run of the comparison between two censuses (with *c*.1-2 million records each) takes 3-5 hours.

3 Delivering data to researchers and the wider public

To provide the highest quality linked data for everyone, it will be made available in two ways for our main target user groups: researchers and the general public. Researchers will be able to access both the historical part (freely downloadable) as well as data protected by the EU-General Data Protection Regulation (GDPR) and its Danish implementation (which offers an additional 10 years of posthumous protection) following the same channels as any other registry-based research. For the public, the website link-lives.dk will display the historical part, unaffected by these protections, where users will be able to search and download limited amounts of data after logging in. By February 2022 we will make available for researchers and the public all life courses for the years 1787 to 1901 from our key sources. Between 2022 and 2024 we will expand the connection to more contemporary sources, so researchers will have access (subject to permission) to multigenerational data for the period 1787-1968.

We have chosen this dual dissemination strategy, instead of publishing open linked data and taking advantage of the semantic web, because of two major constraints. First, a part of the data is proprietary (Ancestry's indexing) and another part contains individuals protected by GDPR/Danish law, so the whole collection could never be freely delivered or downloaded. Second, there is a low penetration of approaches to the semantic web among the users we target, who are still tied to traditional forms of delivery (Excel for historians, csv files or traditional SQL databases for social and health scientists) even though there are promising projects publishing historical databases as open-linked data [31, 32]. Thus, in order to maximize early and high usage of the collection, we have prioritized formats and systems that are already in use, that do not require our users to be trained further. The fit of this decision to the current project has been confirmed through user tests, teaching and other forms of dissemination with the two main user groups – family historians and researchers. As at August 2021 we have received no explicit requests outside of these standard formats but we are aware that there is a growing trend in Digital Humanities in Denmark for working with the Semantic Web. In any case, we could easily move into publishing part of the data as open data in the future, providing that additional funding and collaborations are found.

4 Conclusion

This paper has presented the key aspects of the Link-Lives project, featuring constraints, opportunities, challenges, and results to date. We have shown how a project of this complexity requires competences and expertise from different types of institutions, from archival personnel, historians, and computer scientists. Moreover, a project like this offers archives new ways to engage with research and public audiences and creates the foundation on which to build future collaborative partnerships and funding opportunities. On the methodological side, we have highlighted how it is key to have an understanding of the provenance of historical data, as well as to transparently record the processes it undergoes in metadata form. A clear overview of the ways in which human

and algorithm decisions affect the data are prerequisites for transforming archival holdings into new datasets. Embedding domain experts as key mediators of data ensures that the data's potential and limitations are fully transmitted to the end users.

References

1. Branje, S., Geeraerts, S., de Zeeuw, E.L., Oerlemans, A.M., Koopman-Verhoeff, M.E., Schulz, S., Nelemans, S., Meeus, W., Hartman, C.A., Hillegers, M.H.J., Oldehinkel, A.J., Boomsma, D.I.: Intergenerational transmission: Theoretical and methodological issues and an introduction to four Dutch cohorts. Developmental Cognitive Neuroscience. 45, 100835 (2020). https://doi.org/10.1016/j.dcn.2020.100835.
2. Moreau, C., Bhérer, C., Vézina, H., Jomphe, M., Labuda, D., Excoffier, L.: Deep Human Genealogies Reveal a Selective Advantage to Be on an Expanding Wave Front. Science. 334, 1148–1150 (2011). https://doi.org/10.1126/science.1212880.
3. Vågerö, D., Pinger, P.R., Aronsson, V., van den Berg, G.J.: Paternal grandfather's access to food predicts all-cause and cancer mortality in grandsons. Nature Communications. 9, 5124 (2018). https://doi.org/10.1038/s41467-018-07617-9.
4. Brown, J.R., Enos, R.D., Feigenbaum, J., Mazumder, S.: Childhood cross-ethnic exposure predicts political behavior seven decades later: Evidence from linked administrative data. Science Advances. 7, eabe8432 (2021). https://doi.org/10.1126/sciadv.abe8432.
5. Mandemakers, K., Kok, J.: Dutch Lives. The Historical Sample of the Netherlands (1987−): Development and Research. Historical Life Course Studies. (2020).
6. Digitising Scotland, https://digitisingscotland.ac.uk/, last accessed 2021/02/23.
7. Thorvaldsen, G., Andersen, T., Sommerseth, H.L.: Record Linkage in the Historical Population Register for Norway. In: Bloothooft, G., Christen, P., Mandemakers, K., and Schraagen, M. (eds.) Population Reconstruction. pp. 155–172. Springer, (online) (2015).
8. Swedpop, http://service.re3data.org/repository/r3d100010146, last accessed 2021/04/15.
9. Ruggles, S., Fitch, C., Goeken, R., Hacker, J.D., Helgertz, J., Roberts, E., Sobek, M., Thompson, K., Warren, J.R., Wellington, J.: IPUMS Multigenerational Longitudinal Panel.
10. The Canadian Peoples, https://thecanadianpeoples.com/, last accessed 2021/06/28.
11. Kaplanis, J., Gordon, A., Shor, T., Weissbrod, O., Geiger, D., Wahl, M., Gershovits, M., Markus, B., Sheikh, M., Gymrek, M., Bhatia, G., MacArthur, D.G., Price, A.L., Erlich, Y.: Quantitative analysis of population-scale family trees with millions of relatives. Science. 360, 171–175 (2018). https://doi.org/10.1126/science.aam9309.
12. Thomsen, A.R.: Lykkens smedje?, social mobilitet og social stabilitet over fem generationer i tre jyske landsogne 1750-1850, PhDissertation, University of Copenhagen (2010).
13. Johansen, H.C.: Danish Population History. University Press of Southern Denmark, Odense (2002).
14. Novo Nordisk Fonden: Artificial intelligence will transcribe the family relationships of Danes and strengthen research, https://novonordiskfonden.dk/en/news/kunstig-intelligens-skal-kortlaegge-danskernes-stam-trae-og-styrke-forskning/, last accessed 2021/03/12.

15. Clausen, N.F.: The Danish Demographic Database—Principles and Methods for Cleaning and Standardisation of Data. In: Bloothooft, G., Christen, P., Mandemakers, K., and Schraagen, M. (eds.) Population Reconstruction. pp. 3–22. Springer, (online) (2015).
16. Van Zeeland, N., Gronemann, S.T.: Participatory Archives. In: Benoit, E., III and Eveleigh, A. (eds.) Participatory Archives. pp. 103–114 (2019).
17. Baker, J.L., Olsen, L.W., Andersen, I., Pearson, S., Hansen, B., Sørensen, T.I.: Cohort Profile: The Copenhagen School Health Records Register. Int J Epidemiol. 38, 656–662 (2009). https://doi.org/10.1093/ije/dyn164.
18. Juel, K., Helweg-Larsen, K.: The Danish registers of causes of death. Dan Med Bull. 46, 354–357 (1999).
19. Van Leewen, M., Maas, I., Miles, A.: HISCO. Historical International Standard Classification of Occupations. Leuven University Press., Leuven (2002).
20. WHO: ICD-10, https://icd.who.int/browse10/2016/en, last accessed 2018/12/05.
21. DigDag.dk, http://digdag.dk/, last accessed 2020/03/26.
22. SHiP: Studying the history of Health in Port Cities, https://www.ru.nl/rich/our-research/research-groups/radboud-group-historical-demography-family-history/ship/, last accessed 2021/08/31.
23. Wisselgren, M.J., Edvinsson, S., Berggren, M., Larsson, M.: Testing Methods of Record Linkage on Swedish Censuses. Historical Methods: A Journal of Quantitative and Interdisciplinary History. 47, 138–151 (2014). https://doi.org/10.1080/01615440.2014.913967.
24. Ruggles, S., Fitch, C.A., Roberts, E.: Historical Census Record Linkage. Annual Review of Sociology. 44, null (2018). https://doi.org/10.1146/annurev-soc-073117-041447.
25. Antonie, L., Inwood, K., Lizotte, D.J., Andrew Ross, J.: Tracking people over time in 19th century Canada for longitudinal analysis. Machine Learning; Dordrecht. 95, 129–146 (2014). http://dx.doi.org.ep.fjernadgang.kb.dk/10.1007/s10994-013-5421-0.
26. Christen, V., Groß, A., Fisher, J., Wang, Q., Christen, P., Rahm, E.: Temporal group linkage and evolution analysis for census data, https://openproceedings.org/2017/conf/edbt/paper-269.pdf, (2017). https://doi.org/10.5441/002/EDBT.2017.83.
27. CLARIAH/burgerLinker. CLARIAH (2021).
28. Edvinsson, S., Engberg, E.: A Database for the Future. Major Contributions from 47 Years of Database Development and Research at the Demographic Data Base. Historical Life Course Studies. (2020).
29. Feigenbaum, J.J.: A Machine Learning Approach to Census Record Linking.
30. Abramitzky, R., Mill, R., Pérez, S.: Linking individuals across historical sources: A fully automated approach*. Historical Methods: A Journal of Quantitative and Interdisciplinary History. 1–18 (2019). https://doi.org/10.1080/01615440.2018.1543034.
31. Hoekstra, R., Meroño-Peñuela, A., Dentler, K., Rijpma, A., Zijdeman, R., Zandhuis, I.: An Ecosystem for Linked Humanities Data. In: Sack, H., Rizzo, G., Steinmetz, N., Mladenić, D., Auer, S., and Lange, C. (eds.) The Semantic Web. pp. 425–440. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-47602-5_54.
32. Hooland, S. van: Linked data for libraries, archives and museums: how to clean, link and publish your metadata. Facet Publishing, London, [England (2014).