LINK-LIVES: BUILDING HISTORICAL BIG DATA FROM ARCHIVAL RECORDS FOR USE BY RESEARCHERS AND THE DANISH PUBLIC

Manuscript paper prepared for the workshop "Innovation on new digital exponential technologies towards the generation of Business Models", Alicante, Archivo Histórico Provincial, 2-3rd September, 20221

Bárbara Ana Revuelta-Eugercios
Project senior researcher at the *Danish National Archives* / Research Associate Professor at the *University of Copenhagen*
bre@sa.dk, +45 41 77 74 88

Abstract
The *Link-Lives* project, which is a cross-disciplinary research project, will take the difficult and time-consuming task of combining information from diverse archival sources relating to any given person, to build life-courses and family relations from 1787 to the present and make them freely and easily available. This will, on the one hand, expand the scope of registry-based research from decades to centuries and open up new avenues for intergenerational research in the health and social sciences and, on the other hand, ease the access to some of Denmark's digital treasures to the average citizen. It is a collaboration of the Danish National Archives, the Copenhagen City Archives and the University of Copenhagen. It is funded through two grants by the Innovation Fund Denmark, the Carlsberg Foundation and two small grants by the Ministry of Culture.

1. Introduction: two user-groups underserved by archives in Denmark

There are currently two user groups still underserved by the Danish archives. On the one hand, the community of researchers from the Humanities, the Social Sciences, and even the health sciences, who are interested in the lifecourse and multigenerational mechanisms that affect the biological and social lives of humans. On the other hand, the large pool of citizens who would like to get started in family history but do not have the time or the competences to do it on their own.

Let me illustrate the first case with the story of fictional Hannah, a PhD student in Humanities or Social sciences at a Danish university, who is interested in the new field of research that focuses on analyzing the intergenerational mechanisms explaining social and biological life. She has read, for example, research using 400 years of Canadian historical parish records that showed that there was a selective advantage to moderate fertility in the frontier population. The paper showed that "while high fecundity was associated with a larger number of children, perhaps paradoxically, moderate fecundity maximized the number of descendants after several generations" (Galor and Klemp 2019). She has also read about how the adverse health consequences of a birth out of wedlock in Sweden in the 1920s were not restricted to the child himself but could also be felt in their offspring and their offspring's offspring, the later born in the late 20th century (Modin, Koupil, and Vaguero 2006). Also, how US researchers showed that being a child

in the US in the 1940s and living in areas with a variety of ethnic groups was related to particular political affiliations 7 decades later (Brown et al. 2021).

Impressed with the possibilities of this type of research, she wonders: “Could these studies be done in Denmark?” She knows that the Central Person Registry, which was established in Denmark in 1968, has very high quality registry data. This registry connects all personal information of residents in Denmark through a unique personal identification number, which allows all types of registries to be combined. The challenge is how to reconstruct the lives of generations before 1968. As soon as she starts googling historical sources, the fantastic news for Hannah is that the Danish archives, national, municipal and local hold a wealth of treasures, that could indeed allow reconstructing lifecourses and generations all the way back to the first census of 1787 – for some areas, even before in time, to the first parish registries that started in the 16th century. The even better news is that there are millions of images of these sources already scanned and photographed, many of them freely available, and also of millions and millions of transcribed records.

Indeed, there are individual and household data from the census sheets and the parish records are fully available as digital facsimiles, together with full transcriptions of more than 10 censuses, in the Danish Demographic Databases, the result of a crowdsourcing project that started in 1992 (Clausen 2015). Other sources of images and transcribed data are, for instance, the municipal archives, which have also run their own digitation and crowdsourcing projects making other local sources available. Additionally, genealogical companies, e.g., Ancestry, have also created new images and transcriptions, for instance, of the parish records, which have just made freely available in Denmark. Many research projects have also created historical data registries, especially for health, that contain precious information on the early 20th century. And, last but not least, there is a host of small databases held by genealogical associations, private citizens and other actors that also have digitized and transcribed enormous amounts of data from the archives.

However, while the resources are there and reconstructing lifecourses and generations back in time is entirely possible, the task is way out of Hannah’s reach and her project’s or of any other single researcher or research project for that matter. There is simply too much data in very different formats, qualities, etc., which makes it impossible for her to fully take advantage of this digital treasure trove because, to all intents and purposes, it is constituted of isolated digital islands of information. Hannah may well decide to try to acquire Canadian, Swedish or American data for her research instead.

The second case I want to illustrate is that of Hans, a middle-aged blue-collar worker who has always been interested in family history. After the birth of his grandson, he finally decides to make a family genealogy and he googles “how to get started in family history”. The first thing that shows up is the company MyHeritage, well established in Denmark, and he gets very excited as it all seem very easy. Soon, however, he hits a paywall, abandons My Heritage – as he is not yet willing to pay for it, - and googles some more and starts finding some of the resources that Hannah found: the National archives, the municipal archives, the genealogy pages, forums, Facebook… What is the difference between pages, why are there so many versions of the censuses, … Where does he start? Every link sends him to more and more opportunities and more datasets and images of descriptions or pages… So, finally, it becomes too much for him, there is too much

information, too much to figure out, and he does not have the time. He many decide that maybe when he retires, he will have a look at it, when he will have the time.

2. The *Link-Lives* project

In short, the *Link-Lives* projects aims at solving the needs of these two types of users and in doing that, also at making data available in a format that may attract even more type of users: the general public, educators and students, and other archives and cultural heritage institutions. Link-Lives is a cross-disciplinary research project that will take the difficult and time-consuming task of combining information from diverse archival sources relating to any given person, to build life-courses and family relations from 1787 to the present and make them freely and easily available. This will expand, on the one hand, the scope of registry-based research from decades to centuries, and open up new avenues for intergenerational research in the health and social sciences, and, on the other hand, will ease the access to some of Denmark's digital treasures to the average citizen. There is a large demand of this type of datasets placed on the institutions that have been able to create them. For instance, more than a thousand articles have been written on data from the databases held at CEDAR, which covers some areas of Northern Sweden. More and more researchers are trying to find multigenerational and genealogical data, (the later, even if flawed) to conduct this type of research (Song and Campbell 2017; Ruggles 2014; Kaplanis et al. 2018), and new projects focusing on intergenerational aspects are being funded (genpop n.d.).

In order to create these lifecourses we combine machine learning, historical research, bioinformatics and citizen involvement to transform Danish archival sources into multigenerational big data. This endeavor is only possible through the cooperation of the Danish National Archives (*Rigsarkivet*), Copenhagen City Archives (*Københavns Stadsarkiv*) and the University of Copenhagen and the funding obtained through two generous large grants by the Innovation Fund Denmark and the Carlsberg Foundation and two small research grants by the Ministry of Culture. The timeframe for the project is 2019-2024.

The process of converting archival material into big historical data will be done and stored into what we call *Link-Lives Links*, a central data infrastructure hosted at the National Archives, which will be disseminated to users through two services: *Link-Lives Science*, a service to researchers to facilitate research, and *Citizen*, a public webpage, where everyone will be able to search and explore the links and data created by the project – i.e., data not protected by the GDPR and its Danish implementation,- which in effect means that of individuals born before 1901.

In order to avoid that the result of the project becomes a data cemetery, the setup will make the continuous integration of new data possible, even after the project's end. These will come either from additional sources transcribed by the ongoing crowdsourcing projects at the Danish archives, or from agreements with other genealogical societies. They could also come from integration of new archival records covering different dimensions of individual lives transcribed by research projects, as paid-for service, that will be part of the income-funded activities at the National Archives, where researchers could pay for transcription or linking of their collections to Link-Lives. Thus, the end result is not a

dataset but an exponentially growing infrastructure able to incorporate historical data about individuals in a dynamic manner.

The project has two phases, the period 2019-2022, focusing on publicly available data not protected by the GDPR until 1901, followed by the integration of GDPR protected data in 2022-2024. In phase 1 we develop a proof of concept, i.e., that this type of data integration is possible, by linking three sources that have been created by three different actors, two archives and a genealogy company. We are currently linking the 10 fully transcribed censuses held at the Danish National Archives, from 1787 to 1901. These contain information on every person in the country in each census year, comprising more than 10M records, that have been obtained through crowdsourcing. As I have mentioned, this has been made possible by volunteers who have been transcribing and scanning for more than 20 years (Clausen 2015; 'Dansk Demografisk Database' n.d.). We link censuses to each other, but also link them to the parish records that have been indexed by the genealogical company Ancestry covering the period 1812 to 1911 (Van Zeeland and Gronemann 2019). They include information on baptisms, marriages and burials, and contain more than 22M records. We also link to the Copenhagen City Funeral records, which includes all burials in the city between 1860 and 1940, including the person's cause of death, which are being transcribed also as a part of a crowdsourcing project at the Copenhagen City Archives (Van Zeeland and Gronemann 2019). Starting in 2022, we will extend our coverage to the early 20th century, through additional censuses, and connect to the modern CPR registry.

At the end of our current funding, we hope to secure additional funding that will allow us to extend the collection with richer sources from the same two archives but also those of other cultural heritage institutions. This is possible because there is an enormous amount of transcribed archival records that are already waiting to be included. All these ensures the project's growth path in three ways. First, through the contributions made available by volunteers who have scanned and photographed millions of images and also transcribed and indexed many millions of records. The value of these contributions amounts to millions of Danish kroner. Second, through collaborations with private companies. And third, through the increasing amount of research projects whose data could be included. For instance, we have a collaboration with researchers at the University of Southern Denmark to link a new dataset on biographies of students who graduated high school during the 19th century to the main Link-Lives. Also, there are already plenty of transcribed projects, as the Cause of death Register, which contains information on all death certificates for the country starting in 1942 and extending into the period cover by modern registration (Juel and Helweg-Larsen 1999). Moreover there are recently funded projects also with exciting prospects. For instance, there is a new project hosted at the National Archives, the *Multigenerational Registry* (Novo Nordisk Fonden n.d.), which is looking into creating even more records from parish registries, focusing on the transcription of the parish records of Denmark from 1920 to 1968 through the development of new text recognition technologies.

The main institutions behind the project are the National Archives, Copenhagen City Archives and two departments at the University of Copenhagen. The team we have assembled to carry out of project reflects the wide array of expertise needed to convert historical sources, which start their life as analogue pieces of paper, and ensure multiple transformations through scanning, transcription, standardization and linking until they

become historical big data. We have a combination of archivists, historians and historical demographers in close collaboration with data scientists and biostatisticians that ensure that we develop the best methods, those that fully incorporate an accurate knowledge of historical sources, lives, societies and reflect the latest trends on state-of-the-art entity resolution. We also ensure that archivists and historians can engage and disseminate to students, family historians and the public at large.

Moreover, to ensure that our work aligns and builds on the current state of the art in linking, we collaborate with research teams at foreign universities with large experience in linking historical records and carrying out intergenerational research and developing new methods, in Sweden, Norway, Scotland and the Netherlands. Our colleagues at the University of Umeå (Edvinsson and Engberg 2020) and Tromsø (Thorvaldsen, Andersen, and Sommerseth 2015) have been engaged in creating historical demographic databases from historical records from the late 1980s, so they have ample experience in treating and dealing with Nordic material, which tends to share many similarities. The new project *Digitizing Scotland* at the University of Edinburgh is, on the other side, another relative newcomer which is transcribing and linking all vital registration in Scotland for the period 1850-1950, leaning heavily on developing new methods to deal with millions of records (Akgün et al. 2020). The group from the Radboud University Nijmegen have experience developing and working on a variety of historical datasets from Dutch historical records. (Mandemakers and Kok 2020).

In the following sections I describe with more detail how we are linking data and how we are planning to deliver it through our services: *Science* and *Citizen*.

3. *Link-Lives Links*: creating links and lifecourses

Given that the term “linked data” has different interpretations in data management and dissemination, let me briefly explain what we mean when we talk about linking. We are not talking about the Semantic Web in this project but on the field of entity resolution (Christen 2012). A “link” for us is the relationship between two records containing personal information that come from two historical different sources, like two census records from two years that we think belong to the same person. By chaining links from different sources, we can create lifecourses that give a reconstructed image about the events that individuals went through in their lives and, from the contextual information in them, reconstruct also family and kinship relations and reconstruct generations.

While on the outset the process seems very similar to that of genealogy and, in a sense, it is, the way we decide if something is a link is different. Where they cross-check different sources from different types of media following a single individual, we need to implement linking methods that can be scaled up to millions of records, which, for now, leaves us at the level of pair-wise linking, i.e., between records in two sources, which is the standard in the literature in entity resolution now. We use the expression “the most probable link” to acknowledge the fact that we can never know if a link was right or no, as there is no way to go back and check, there is no real “ground truth”, as it is called in machine learning. Instead, we develop different methodologies that allow us to arrive to our best estimate of whether any two records are a link. And we develop those methods with maximum focus on transparency, reliability and reproducibility, ensuring each link as metadata, so anyone can assess, reproduce or challenge our methods.

In this project we implement three methods of linking. First, we create sets of linked records through manual linking. That means that a human (a domain expert/historian) takes a decision on whether the information from two records fulfills the conditions to be considered a link. We have created a software, *Assisted Linking Application* (ALA), that enables computer-assisted linking and developed a set of protocols and guidelines to ensure systematic data creation. We use two independent linkers whose disagreements are afterwards resolved by an arbiter. As of August 2021, we have created more than 35.000 records for different types of sources, year ranges and areas. The construction of this data is a cornerstone of our work because, in the absence of true ground truth, it is what allows us to have a best estimate of what a domain expert thinks is a link. And this data is then key to test and train automatic models. Our data shows that humans can find up to 80% of cases in most instances, but there is a large variation between geographies, chronologies, sources and users.

The second type of method is a set of rule-based algorithms where a historian and a data scientists program a set of rules that can be implemented across the whole dataset. These have been widely spread in many projects as they are relatively easy to implement (Ruggles 2002; Ruggles, Fitch, and Roberts 2018; Thorvaldsen, Andersen, and Sommerseth 2015; Fu et al. 2014). A simple rule could be that two records need to have a close enough name, place of birth and age to be a link. Of course, the devil is in the details, e.g., how close do they need to be to be considered “close”? Comparing the model results with our domain-expert created data, models capture around 70% of what human domain experts can and the whole linkage program can be run in 2-3 hours in our high-performing environment, while it takes around 3-4 person hours to fully link, compare and resolve 100 records of domain expert data.

The third type of models is comprised of other automatic methods and machine learning approaches. In the simplest of terms, in machine learning, we take the small set of records created by our domain experts (called “training data”) and feed them to a model that, then, figures out from that data what is a link and what is not a link and that can produce prediction, also for the whole dataset. And if we save part of the linked data and do not use it completely to train the model, we can then use it to test how the model compares to our domain experts. We have recently gathered enough and varied-enough training data and are in the process of testing different implementations. We are implementing methods already in use in the literature in order to benchmark and compare them, including the Expectation-Maximization approach (Abramitzky, Mill, and Pérez 2019), support-vector machine (Ruggles et al., n.d.; 2011; Antonie et al. 2014) but also testing other models, e.g., random forest and variation recurrent neural networks. Their results are very encouraging. Each of these methods have their pros and cons, so they can be used for different purposes by our users, depending on their interests.

However, what it is clear is that the linked intergenerational data researchers like our user Hannah would get from Link-Lives will reflect the historical reality where it originated, but it will have travelled a very long way and experienced substantial transformations that need to be considered. And this is also central part of *Link-Lives*’ mission, to ensure that we document and highlight each step of these explicit or implicit human, computing or human and computing decisions that is responsible for the final data: reality was captured because of government decisions; the registration was implemented by statistical offices and individual agents; individuals may have different levels of willingness to accurately

report about their lives; archives may have had chronologically and geographically different preservation policies; archives may also have radically different digitation policies; the aims and initiators of crowdsourcing projects affect the design of what and how information was captured, the volunteers may have different levels of competences and willingness to follow the rules; the *Link-Lives* team has taken many decisions to standardize, process and link data as well as dissemination formats.

This complex lifecourse does not mean that the data is not usable or high quality but that, in order to obtain the highest quality research, we need to make available sufficient documentation capturing these different steps, including metadata. Our aim is that our users can always distinguish between our different levels of interpretation and choose what fits them better.

4. *Link-Lives Citizen* and *Science*: delivering data in the format that users need it

Given our understanding that our users will have different interests, profiles, and competences we have designed a dissemination strategy that take these into account.

LINK-LIVES CITIZEN

To cater for the genealogy and volunteer community, we have designed a search function in our webpage where anyone can freely access all our created links and lifecourses. The data from Links that can be made available for the public because it is not protected by the data privacy legislation, which in its Danish implementation protects individual data up to 10 years after their death. We are in the final stages of the construction of the front and backend and we expect to launch by early 2022. As of now, there is only information about the project in our webpage, linklives.dk, but soon it will be possible to do simple and advanced searches on individuals and be able to retrieve the results generated by our different approaches. After searching, users will be able to scroll through both our original sources and re-created lifecourses, which they will be able to explore. The main difference from our page from other types of similar genealogy resources is that we do explicitly present the users with the methods employed for generating every single link in the form of easily accessible metadata. It is very important for us to show that we do not aim to declare who is someone's great-grandmother or attest that these are "real" lifecourses, but just provide different options for users that make it easier for them to find what they are looking for.

A second difference, important for us, is the inclusion of a feedback function that will allow us to both deal with feedback in a structured manner and gather additional data that we could use to further refine our methods. Users will need to be logged in and provide feedback clicking some boxes, answering whether the link is correct and why they think it is correct (from the sources already present in Link-Lives, from other sources not yet available for us, or from their own research). These manually verified-links from volunteers and interested family historians will create data very different from our domain-expert, guidelines-constrained linked data. However, we believe that they could help us gather data on the multiple-source-checking way of linking employed by genealogists and, when collected with the right metadata and adequately aggregated, can help us understand more on how pairwise linking compares to multiple-source linking and how to improve our models.

In all the process from design to implementation, we have included several rounds of user-testing, not only with family historians and volunteers, but also with researchers and other general public groups, to ensure that we incorporate user feedback to maximize both functionality and ease.

LINK-LIVES SCIENCE

Link-Lives Citizen will serve as a window for researchers to get a first glimpse of the data but they will actually get access through what we call *Link-Lives Science*, where researchers will be able to download all, fully available, data files for sources, links, lifecourses, metadata and documentation. This service will be hosted at the National Archives and will be operational in February 2022. It will provide all data as a simple service, and any researcher with some programing competences will be able to work with it. It will also include annual releases of new linked data until 2024. However, we hope to secure additional funding to develop an easier interface so that also less programming-savvy researchers, students and the public can also engage with the data in different ways. When we include data protected by GDPR protections and its Danish implementation, access will require the same type of permits and secure access as any other data held and requested at the National Archives.

As part of the development of *Science*, the project itself, and with collaborators, is engaging in research with the data as we develop it. This is a way to ensure that the data is tested as it is developed and that the new insights gathered by the research that we perform provides new knowledge about the source, the data or Danish history that can itself be incorporated into the development. The research that is being developed by us and our collaborators touches a variety of disciplines, from history, historical demography, history of medicine, economic history, onomastics, data science, bioinformatics, archival studies… thus, while this project may look like an infrastructure project, its construction is driven by research in historical methods and other disciplines, where most of the team members are engaged in research in one of the 10 articles that we currently have ongoing: I am a senior researcher myself and two professors, two senior researchers, two postdocs and three PhD students,

5. Perspectives: value and beneficiaries

Overall, we believe that there are main two beneficiaries of the projects are; first, the Danish general public/archives, but also the international research community, who will be able to access new data and will open up unprecedented avenues of research. It will be possible not only to do historical research but also to expand the possibilities in sociology, political science, economics, health sciences and data science. All of this will, at the same time, hopefully attract researchers and research funding to Denmark.

Second, the Danish community of family historians and the general public, who will be able to access freely all that public data in a new form. This will be a way of giving back to the community of volunteers who have transcribed the data through crowdsourcing. We expect that the page will become for many a first “go-to” place for getting started in family history, facilitating the way into the rich ecosystem of resources for genealogy.

Third, for the archives partnering in the project, creating this tool serves as a new way of exploring disseminating their collection and engaging their users. The project provides an opportunity to experiment with a new role of data creation, substantially different from the traditional roles of preservation and dissemination of data. For example, for the National Archives, *Link-Lives* underpins one of the new strategic directions focused on making data available in new ways. For Copenhagen City Archives, *Link-Lives* ensures that the volunteer-research loop mediated by the archive is closed: the engagement of their users in their crowdsourcing projects can be fruitfully used for research, which is later made available for them in new ways, that can lead to new ways of engagement.

Moreover, although it is clear that it is beyond the scope of the project as it is right now, the type of structure we propose opens the door for including other types of digital treasures from other GLAMs (Galleries, Libraries, Archive and Museums) for research or wide-public interest: new archival collections from crowdsourcing in municipal archives, both in the form of traditional registries or in the growing number of letters and other natural-language sources that are becoming available, artists' lifecourses could be connected to their artworks in Danish museums, writers with their publication at the Royal Library and even local personalities to their contributions or artifacts in local museums or archives.

Finally, while there are projects in the US and Europe carrying out some of the elements that are included in this project, i.e., research projects linking large-scale data, archives engaging in transcription, collaboration with genealogical companies, etc., we think that this project has two differentiating features that make it very strong: first, while we in Denmark are relative newcomers to the business of creating historical databases, which has a very long tradition other university departments in Europe, as our colleagues in Norway and Sweden, the enormous wealth of pre-existing data and the new technological developments, have allowed us to put together a project that basically has progressed from nothing to a full nation-wide population database for the 19th century in three years, becoming/arriving among the first countries in the world to be able to do that. Second, our large-scale established collaboration has allowed us to build a business model that tries to reach different types of users and engage into synergies that, by design, will contribute to propel it further into the future beyond the end of our current funding.

6. References

Abramitzky, Ran, Roy Mill, and Santiago Pérez. 2019. 'Linking Individuals across Historical Sources: A Fully Automated Approach*'. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 1–18. https://doi.org/10.1080/01615440.2018.1543034.

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 (2): 130–46. 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; Dordrecht* 95 (1): 129–46. http://dx.doi.org.ep.fjernadgang.kb.dk/10.1007/s10994-013-5421-0.

Brown, Jacob R., Ryan D. Enos, James Feigenbaum, and Soumyajit Mazumder. 2021. 'Childhood Cross-Ethnic Exposure Predicts Political Behavior Seven Decades Later: Evidence from

Linked Administrative Data'. *Science Advances* 7 (24): eabe8432. https://doi.org/10.1126/sciadv.abe8432.
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 Bloothooft, Peter Christen, Kees Mandemakers, and Marijn Schraagen, 3–22. (online): Springer.
'Dansk Demografisk Database'. n.d. Accessed 22 December 2017. http://www.ddd.dda.dk/.
Fu, Zhichun, H. M. Boot, Peter Christen, and Jun Zhou. 2014. 'Automatic Record Linkage of Individuals and Households in Historical Census Data'. *International Journal of Humanities and Arts Computing* 8 (2): 204–225. https://doi.org/10.3366/ijhac.2014.0130.
Galor, Oded, and Marc Klemp. 2019. 'Human Genealogy Reveals a Selective Advantage to Moderate Fecundity'. *Nature Ecology & Evolution* 3 (5): 853–57. https://doi.org/10.1038/s41559-019-0846-x.
genpop. n.d. 'Genes, Genealogies and the Evolution of Demographic Change and Social Inequality'. GENPOP. Accessed 30 March 2021. http://genpop.org/.
Juel, K., and K. Helweg-Larsen. 1999. 'The Danish Registers of Causes of Death'. *Danish Medical Bulletin* 46 (4): 354–57.
Kaplanis, Joanna, Assaf Gordon, Tal Shor, Omer Weissbrod, Dan Geiger, Mary Wahl, Michael Gershovits, et al. 2018. 'Quantitative Analysis of Population-Scale Family Trees with Millions of Relatives'. *Science* 360 (6385): 171–75. https://doi.org/10.1126/science.aam9309.
Mandemakers, Kees, and Jan Kok. 2020. 'Dutch Lives. The Historical Sample of the Netherlands (1987−): Development and Research'. *Historical Life Course Studies*, June. /article/dutch-lives-historical-sample-netherlands-1987%E2%88%92-development-and-research.
Modin, Bitte, Ilona Koupil, and Denny Vaguero. 2006. 'The Impact of Early Twentieth Century Illegitimacy across Three Generations'. Centre for Health Equity Studies, CHESS, Stockholm University.
Novo Nordisk Fonden. n.d. 'Artificial Intelligence Will Transcribe the Family Relationships of Danes and Strengthen Research'. *Novo Nordisk Fonden* (blog). Accessed 12 March 2021. https://novonordiskfonden.dk/en/news/kunstig-intelligens-skal-kortlaegge-danskernes-stam-trae-og-styrke-forskning/.
Ruggles, Steven. 2002. 'Linking Historial Censuses: A New Approach'. *History & Computing* 14 (1/2): 213–24.
———. 2014. 'Big Microdata for Population Research'. *Demography* 51 (1): 287–97. https://doi.org/10.1007/s13524-013-0240-2.
Ruggles, Steven, Catherine A. Fitch, and Evan Roberts. 2018. 'Historical Census Record Linkage'. *Annual Review of Sociology* 44 (1): null. https://doi.org/10.1146/annurev-soc-073117-041447.
Ruggles, Steven, Catherine Fitch, Ron Goeken, J David Hacker, Jonas Helgertz, Evan Roberts, Matt Sobek, Kelly Thompson, John Robert Warren, and Jacob Wellington. n.d. 'IPUMS Multigenerational Longitudinal Panel'.
Ruggles, Steven, Evan Roberts, Sula Sarkar, and Matthew Sobek. 2011. 'The North Atlantic Population Project: Progress and Prospects'. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 44 (1): 1–6. https://doi.org/10.1080/01615440.2010.515377.
Song, Xi, and Cameron D. Campbell. 2017. 'Genealogical Microdata and Their Significance for Social Science'. *Annual Review of Sociology* 43 (1): 75–99. https://doi.org/10.1146/annurev-soc-073014-112157.
Thorvaldsen, Gunnar, Trygve Andersen, and Hilde L. Sommerseth. 2015. 'Record Linkage in the Historical Population Register for Norway'. In *Population Reconstruction*, edited by Gerrit Bloothooft, Peter Christen, Kees Mandemakers, and Marijn Schraagen, 155–72. (online): Springer.

Van Zeeland, Nelleke, and Signe Trolle Gronemann. 2019. ‘Participatory Archives’. In *Participatory Archives*, edited by Edward Benoit III and Alexandra Eveleigh, 103–14. http://ebookcentral.proquest.com/lib/kbdk/detail.action?docID=6031675.