How to cite this zine:
Abouchacra, Zeina. (2024). Unlinking the Chain: A Guide to Breaking Data Linkages and Protecting Your Privacy. Hamilton

(PNG tree, n.d.)

Created for:
CNMCS 720: Data Cultures. Department of Communication Studies and Media Arts. Winter 2024. Dr. Andrea Zeffiro

Back Cover

# Unlinking the Chain: A Guide to Breaking Data Linkages and Protecting Your Privacy

By: Zeina Abouchacra

Front Cover

B.C. health-care workers' private information subject to data breach

Pixel Hunt

Facebook Is Receiving Sensitive Medical Information from Hospital Websites

I Just Got a COVID-19 Test. Who Now Knows I Got It?

Data linkage, or record linkage, is all about connecting information from different datasets to get a better understanding of people, events, or entities. It's like putting together puzzle pieces from different boxes to see the full picture. This process involves finding and joining information that belong to the same record across different sources, like databases, administrative systems, or registries. Thanks to advancements in technology, researchers, policymakers, and analysts can now combine information from these different sources to gain deeper insights, make better decisions, and tackle complex research questions.

Despite the many benefits of data linkage, the common discourse about this topic in media focuses on concerns about privacy, data security, and the possible misuse of personal information. News articles tend to shine a light on cases where personal data is compromised, accessed without permission, or exploited. These stories are often about big tech companies like Facebook, TikTok, and Google, raising concerns about how they handle user data and its implications for privacy.

All the Data Amazon's Ring Cameras Collect About You

The popular security devices are tracking (and sharing) more than you might think.

FORBES > INNOVATION > CYBERSECURITY

Facebook's New Link History Update Exposes Browser Risk

Medical-record software companies are selling your health data

CAN WE TRACK COVID-19 AND PROTECT PRIVACY AT THE SAME TIME?

Exposed

The erosion of privacy in the Internet era

Will Google's and Apple's COVID Tracking Plan Protect Privacy?

Is a breakdown in trust, transparency and social cohesion a price worth paying for more extensive data linkage?

So you gave personal info to a company caught in a data breach. Now what?

The media often discusses data linkage, emphasizing the importance of individuals being proactive in protecting their privacy. News articles advise individuals to be careful when sharing personal information online and encourage them to consider using privacy-enhancing tools. While it's essential to be mindful of privacy risks, the media's intense focus on alarming stories can sometimes lead to increased fear and distrust of digital technologies and online platforms.

As brands test Amazon's direct link between digital ads and Whole Foods purchases, they spot new data nuggets — and gaps

iPhone keeps record of everywhere you go

Privacy fears raised as researchers reveal file on iPhone that stores location coordinates and timestamps of owner's movements

Canada's Broken Electronic Medical Records Model

Across BC and the country, patients, doctors and the health-care system suffer from a faulty patchwork of incompatible systems.

TECH / AMAZON / AMAZON ALEXA

Researchers find Amazon uses Alexa voice data to target you with ads

HOME > SECURITY

Does Your Phone Listen to You for Ads? Or Is It Just Coincidence?

Your phone has a built-in microphone. Is your phone listening to you and sharing your secrets with Google or Facebook?

A Phone Carrier That Doesn't Track Your Browsing or Location

The new Pretty Good Phone Privacy service for Android hides the data linking you to your mobile device.

The Government has built a data colossus - is it playing with fire?

# Evolution of Data Linkage

1796

The first instance of data linking is traced back to Edward Jenner, an English physician and scientist who conducted ground-breaking research by studying individuals who had been infected with cowpox and its immunity against smallpox (Smith & Flack, 2021).

1800

John Snow, an English physician, mapped the locations of cholera cases and water pumps in London to identify contaminated water as the source of the outbreak (Moriyama, 1964). His pioneering epidemiological investigation laid the groundwork for future use of data linking in public health.

1880

Herman Hollerith's invention of the tabulating machine in 1880 marked a pivotal moment in data processing history. This machine automated the tabulation of census data, streamlining data management for government agencies and commercial enterprises. Hollerith's invention laid the groundwork for future data linkage advancements (Priestley, 2011).

1935

John Snow, an English physician, mapped the locations of cholera cases and water pumps in London to identify contaminated water as the source of the outbreak (Moriyama, 1964). His pioneering epidemiological investigation laid the groundwork for future use of data linking in public health.

1959

Herman Hollerith's invention of the tabulating machine in 1880 marked a pivotal moment in data processing history. This machine automated the tabulation of census data, streamlining data management for government agencies and commercial enterprises. Hollerith's invention laid the groundwork for future data linkage advancements (Priestley, 2011).

1964

Sir Donald Acheson founded the Oxford Record Linkage Study, which connected birth, morbidity, and mortality data for an entire community. This system revolutionized epidemiological research by enabling the analysis of disease patterns over time (Acheson, 1964).

1969

Fellegi and Sunter's (1969) research laid the groundwork for probabilistic data linkage, offering a theoretical foundation for record linkage methodologies.

1990

The adoption of electronic health records (EHRs) in the 1990s revolutionized healthcare data management, enabling more efficient storage, retrieval, and sharing of clinical information among healthcare providers (Shah et al., 2010).

2000

Governments accelerated the diffusion of EHRs, promoting the development of interoperable systems to facilitate data exchange and translational medical research. These electronic patient databases, became important tools for healthcare providers to input, store, and retrieve patient data, supporting the integration of data for research purposes

2010

Privacy-preserving record linkage techniques, such as the use of Bloom filters, emerged to protect personal information during data linkage processes, especially in the healthcare sector (Christen, 2019).

2020

The COVID-19 pandemic in 2020 demonstrated the importance of data linkage in public health. Data linkage initiatives played a crucial role in tracking the spread of the virus, verifying vaccinations, monitoring healthcare capacity, and informing evidence-based decision-making to mitigate the impact of the pandemic

# How is Data Linked?

Two primary methodologies guide the data linkage process: deterministic and probabilistic linkage. These methods serve as the cornerstone for connecting disparate datasets, each offering distinct approaches to matching records.

## 1. Deterministic Linkage

Deterministic linkage is a method used to match records from different datasets by requiring an exact match on specific identifiers or attributes (Shah et al., 2010). It's like having a strict checklist mandating that both records must meet to be considered a match.

(Slides Carnival, n.d.)

For example, if two records have the same social security number, date of birth, and name, they are deemed a match. This method leaves no room for error - if there's any discrepancy, even a small one, the records are not linked together. Deterministic linkage assumes that there are no errors in the data and follows a rigid approach where any deviation from the checklist results in records being labeled as non-matches.

## 2. Probabilistic Linkage

Probabilistic linkage takes a more flexible approach. Instead of demanding an exact match, it assesses the likelihood of a match based on similarities between records (Asher et al., 2020). Even if records don't match perfectly, probabilistic linkage assigns a probability score to pairs of records, indicating the likelihood of being a true match.

This method considers errors in matching variables and evaluates various attributes to calculate the probability of a match based on similarities. It's like a scale that measures how likely it is for two records to be a match, considering differences in the data rather than demanding exact matches.

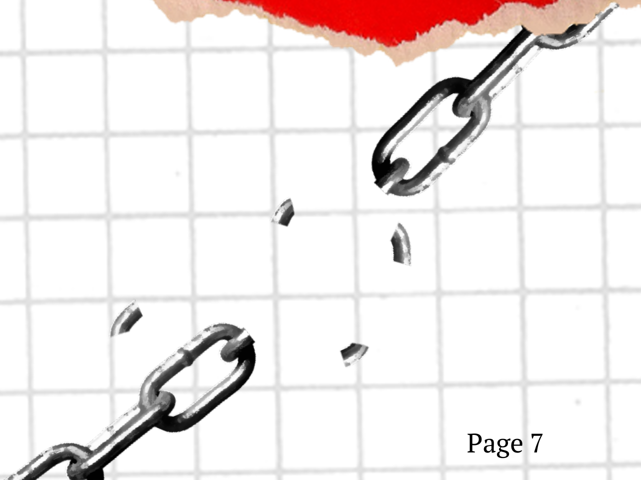
# Case Study 1: Healthcare

## Limiting Infectious Disease Outbreaks

Data linkage is crucial for global public health efforts, especially during infectious disease outbreaks like the COVID-19 pandemic. It helps authorities understand various aspects of the virus and guides response strategies effectively.

One major advantage of data linkage is its ability to offer a comprehensive view of disease dynamics, including how diseases spread, the factors that contribute to transmission, and the outcomes for affected individuals (Field et al., 2023). This broader perspective empowers public health authorities to create targeted interventions and allocate resources more efficiently to curb the spread of infectious diseases (Field et al., 2023).

Data linkage plays an important role in evaluating preventive measures, such as vaccination programs. By tracking vaccine uptake and effectiveness across different populations, linked data helps optimize immunization strategies and ensures widespread protection against infectious diseases. Additionally, this approach can help pinpoint gaps in healthcare utilization and access, allowing authorities to address disparities and enhance healthcare delivery to vulnerable communities.



## Providing Individual Healthcare Delivery

Data linkage is also used in health care with administrative data to create a comprehensive understanding of an individual's health service journey by integrating information from various administrative data sources (De Oliveira et al., 2022). These sources include hospitalizations, emergency department visits, primary care consultations, medication dispensing, and psychiatric services.

For example, consider a patient who visits the emergency department for a broken arm. Through data linkage, this visit can be connected to other healthcare interactions such as subsequent hospital stays, visits to specialists, or appointments for diagnostic tests. Analyzing this linked data allows healthcare providers to track the patient's health progress over time, identify trends in their healthcare usage, assess the effectiveness of different treatments, and ensure continuity of care across different healthcare settings (De Oliveira et al., 2022).

Administrative health data linkage also helps identify underlying health conditions and risk factors associated with specific diseases. By linking hospitalization records with medication data, healthcare providers can identify chronic conditions like diabetes or high blood pressure and monitor their management and outcomes. This comprehensive view of a patient's healthcare encounters supports better risk assessment and personalized care planning.

## Improving Health Care Research

Routinely collected healthcare data, sourced from disease registries, primary and secondary care databases, administrative records, and public health reports, serve as essential resources for healthcare professionals to undertake pioneering research endeavors. By linking these data sets, researchers and health care professionals can match groups of individuals, uncovering valuable insights into disease associations that are otherwise difficult to investigate (through traditional research methods such as randomized controlled trials) (Jutte et al., 2011). This has enabled investigations into connections between conditions like gall bladder disease and colon cancer, appendectomy and inflammatory bowel disease, and vasectomy and prostate disease (Jutte et al., 2011).

Data linkage has also facilitated the exploration of multiple and overlapping outcome domains within the same group of individuals. For example, studies have assessed both medical outcomes (such as hospitalization rates and mortality) and educational outcomes (like academic performance) in children from population cohorts. This comprehensive approach allows researchers to explore various facets of health and well-being.

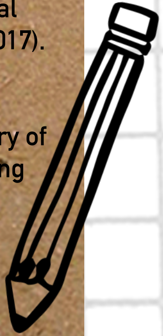
The use of data linkage has proven to be invaluable in population-based prediction research. For instance, in Ontario, Canada, researchers devised an algorithm called the Diabetes Population Risk Tool (DPoRT), which accurately predicts diabetes risk at a population level using self-reported measures gathered from routine health surveys (Rosella et al., 2010). This method of estimating disease incidence facilitates more efficient population health planning and allows for the evaluation of the effectiveness of illness prevention strategies, ultimately contributing to improved public health outcomes.

# Case Study 2: Government

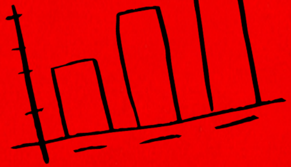
## Supporting Policy Making

The linkage of records across different databases has become a powerful tool for governments to make informed decisions, allocate resources effectively, and address complex societal challenges for the benefit of citizens and communities.

For instance, in Ontario, Canada the linkage of administrative health care databases with data from Immigration, Refugees and Citizenship, Canada's permanent resident registry, the Office of the Registrar General's Vital Statistics Death Registry, and the federal Indian Register has yielded valuable insights (Guttmann et al., 2017). This has allowed governments to understand health services utilization across most healthcare sectors, including hospital, outpatient, emergency, and long-term care as well as the delivery of health care services among different immigrant classes (including economic immigrants, family class immigrants, and refugee or asylum seekers (Guttmann et al., 2017).



The collaboration between the Ontario Ministry of Children, Community and Social Services, responsible for administering social assistance programs, and organizations like ICES exemplifies the utility of data linkage in meeting the diverse needs citizens and creating positive social impact (Guttmann et al., 2017). By linking data across different data basis and partner organizations, government agencies can now enhance their decision-making, policy development, and service provision.



## Improving Population Reporting

Census data, collected periodically, provide crucial information about a country's population, including demographic, social, and economic characteristics However, traditional census records have limitations in capturing dynamic demographic processes and addressing complex research questions. To overcome these limitations, governments worldwide have increasingly embracing data linkage techniques to enhance the accuracy, completeness, and utility of census data.

For example, after the hurricanes Laura and Delta impacted United States Census Board ability to count non-responding households in Louisiana (including areas in Allen, Beauregard, Calcasieu, and Jefferson Davis parishes) (Mule, 2021). The Bureau used data from the Medicare Enrollment Database, Indian Health Service Patient Registration, Selective Service System Registration, and past census survey data to enumerate households in the affected regions (Mule, 2021). The reliance on linked administrative records allowed the Bureau to make more efficient and complete census and played an important role in understanding long-term societal changes informing contemporary policies

Administrative data offer high-quality demographic information that can complement census data and help fill in gaps or correct inaccuracies. By linking administrative records to census responses using personally identifying information (PII), governments can enhance the completeness and accuracy of census datasets.

## Ensuring Road Safety and Security

Various sources, including police reports, hospital records, and mortality data from coroner systems, contribute to traffic safety data. However, each data source has its limitations. For instance, hospital records offer detailed information about sustained injuries but lack information about car accidents and roadway characteristics. While police or insurance reports provide extensive details on car accident but lack data on the severity of injuries.

To address these challenges, governments utilize data linkage strategies to connect crash data with medical records. For example, in the United States, initiatives like Maryland's Crash Outcome Data Evaluation System (CODES) employs probabilistic methods to link various datasets, including those from police, EMS, hospitals, and death certificates (Smith, 2015). This linked data has been instrumental in conducting a wide range of studies, such as assessing the effectiveness of seat belts, analyzing patterns of injuries in different types of collisions, examining the impact of newer vehicles on safety, and studying the effects of external factors like casino gambling on alcohol-related crashes (Smith, 2015).

By linking diverse traffic related datasets, governments gain valuable insights into the causes and consequences of traffic incidents, allowing them to develop targeted interventions and policies to improve road safety. These data linkage efforts enable authorities to identify high-risk areas, evaluate the effectiveness of existing safety measures, and implement evidence-based strategies to prevent injuries and save lives on the roads.

# Case Study 3: Private Sector

## Personalizing Online Advertisements

Data linkage serves as a fundamental tool in delivering personalized experiences to clients through targeted advertising campaigns. By integrating data from diverse sources such as online interactions, in-store purchases, and social media engagement, marketers gain a comprehensive understanding of consumer behavior and preferences. This insight allows tailored advertisements and product recommendations to be shared with specific audience segments, thereby enhancing the overall customer experience.

For instance, Experian Marketing Services, a provider of data-driven marketing, has a platform called OmniView which offers marketers and advertisers a single customer view by establishing identification keys for consumers across different touchpoints (including social, email, mobile, and transactional data) (Experian Marketing Services, 2014). This integrated approach allows marketers to create detailed customer profiles and deliver personalized advertising messages based on individual preferences and past interactions.

By leveraging data linkage, marketers can deliver advertisements that are tailored to individual interests and preferences, making the overall advertising experience more enjoyable and engaging for consumers. For example, someone who enjoys outdoor activities may receive ads for hiking gear or camping equipment, while someone interested in fashion may receive ads for clothing brands they are likely to enjoy. This personalized approach not only enhances the consumer experience but also increases the likelihood of discovering products or services that meet their needs and interests. .

## Tailoring Financial Services

Data linkage plays a role in the financial services sector, particularly in providing personalized banking solutions to customers. Banks and financial institutions utilize data from various sources such as transaction history, spending patterns, credit scores, and demographic information to understand their customers' financial behavior and preferences (Baker & Kueng, 2022). By linking this data, financial institutions can create comprehensive customer profiles and offer tailored banking products and services.

For example, fintech companies, banks, investment firms, and asset management companies aggregate data from multiple financial accounts, including bank accounts, credit cards, loans, and investments, to provide users with a holistic view of their financial health. By analyzing spending habits and financial goals, these platforms can offer personalized recommendations for budgeting, saving, investing, and managing debt. Data linkage also enables these financial institutions to offer personalized loan and mortgage products with tailored interest rates and repayment terms based on the customer's financial history and risk profile (Baker & Kueng, 2022).

Through data linkage, financial institutions can also enhance fraud detection and risk management capabilities. By analyzing transaction data and detecting unusual spending patterns or suspicious activities, banks can quickly identify and mitigate fraudulent transactions, protecting both customers and the financial institution.

## Offering Individualized Product Recommendations

In the e-commerce sector, data linkage is instrumental in providing personalized product recommendations to customers and enhancing their shopping experience. E-commerce platforms collect and analyze data from various sources, including browsing history, purchase behavior, product reviews, and demographic information, to understand customer preferences and interests (Helms et al., 2008). Using sophisticated algorithms and machine learning techniques, e-commerce platforms can then generate personalized recommendations tailored to each customer's unique profile.

For example, Amazon utilizes data linkage to power its recommendation engine, which analyzes customers' past purchases, browsing history, and interactions with the platform to deliver personalized product recommendations and targeted promotions in real-time. This helps individuals save time and effort that would otherwise be spent searching through numerous products to find what they are looking for. Additionally, personalized recommendations increase the likelihood of customers finding products that meet their specific needs, resulting in higher satisfaction with their purchase decisions.

# Is My Data Linked?

Take the quiz below to see if you're safe from data linkage. Read each question carefully and check off the response that best applies.

Do you use the Internet?

○ Y es ○ No

Have you removed or erased the data from your computer or cloud service?

○ Yes ○ No

Did you go to a doctor or dentist in the past year?

○ Yes ○ No

Have you used your phone or computer to do a google search in the past 30 days?

○ Yes ○ No

Have you liked a post on social media this week?

○ Yes ○ No

Have you paid any bills online?

○ Yes ○ No

Do you search the internet using a private browser/ incognito mode?

○ Yes ○ No

Have you purchased something online this year?

○ Yes ○ No

Do you have a credit card you use?

○ Yes ○ No

Have you ever participated in a national census?

○ Yes ○ No

Do you have a social insurance number?

○ Yes ○ No

Do you anonymize your personal information online?

○ Yes ○ No

Do you participate in loyalty programs while shopping?

○ Yes ○ No

Does your phone have location services turned on?

○ Yes ○ No

Have you picked up a prescription from a pharmacy in the past 5 years?

○ Yes ○ No

# Quiz Results

Tally up how many times you checked off yes and no to learn more about your results

3-5
No

Making sure your data is not linked seems to be fairly important to you, but it's not the focus of your existence. You've tried some techniques to limit how your data is shared, accessed, and used but it's still linked. Perhaps living off the grid might help you stop your data from being linked.

6-8
Yes

While you may have done things manually in the past, you are now getting on board with the technology train. You don't limit your day-to-day interactions out of the fear that your data may be linked. Instead, you are worried about more pressing concerns like passwords being leaked or your computer over heating.

10+
Yes

You live your life without the fear that your data might be linked and enjoy the benefits that this technology affords. You love receiving product recommendations based on your purchase history. You like how your family doctor knows about your most recent visit to the walk-in clinic without having to explain everything to them.

Acheson, E. D. (1964). Oxford record linkage study: A central file of morbidity and mortality records for a pilot population. *Journal of Epidemiology & Community Health*, *18*(1), 8-13. https://doi.org/10.1136/jech.18.1.8

Asher, J., Resnick, D., Brite, J., Brackbill, R., & Cone, J. (2020). An introduction to probabilistic record linkage with a focus on linkage processing for WTC registries. *International Journal of Environmental Research and Public Health*, *17*(18), 1-16. https://doi.org/10.3390/ijerph17186937

Baker, S. R., & Kueng, L. (2022). Household financial transaction data. *Annual Review of Economics*, *14*(1), 47-67. https://doi.org/10.1146/annurev-economics-051520-023646

Christen, P. (2019). Data linkage: The big picture. *Harvard Data Science Review*, *1*(2), 1-6. https://doi.org/10.1162/99608f92.84deb5c4

De Oliveira, C., Gatov, E., Rosella, L., Chen, S., Strauss, R., Azimaee, M., Paterno, E., Guttmann, A., Chong, N., Ionescu, P., Ji, S., Kopp, A., Lan, A., Ma, C., Pring, M., Raj, P., Ryan, S., Saskin, R., & Wong, F. (2022). Describing the linkage between administrative social assistance and health care databases in Ontario, Canada. *International Journal of Population Data Science*, *7*(1). https://doi.org/10.23889/ijpds.v7i1.1689

Dunn, H. L. (1946). Record linkage. *American Journal of Public Health and the Nations Health*, *36*(12), 1412-1416. https://doi.org/10.2105/ajph.36.12.1412

Experian Marketing Services. (2014, September 24). *Experian marketing services launches identity-linkage engine for digital advertising industry to resolve data-quality and accuracy challenges*. PR Newswire: press release distribution, targeting, monitoring and marketing. Retrieved April 5, 2024, from https://www.prnewswire.com/news-releases/experian-marketing-services-launches-identity-linkage-engine-for-digital-advertising-industry-to-resolve-data-quality-and-accuracy-challenges-276974401.html

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183-1210. https://doi.org/10.1080/01621459.1969.10501049

Field, E., Strathearn, M., Boyd-Skinner, C., & Dyda, A. (2023). Usefulness of linked data for infectious disease events: A systematic review. *Epidemiology and Infection*, *151*. https://doi.org/10.1017/s0950268823000316

Guttmann, A., Chiu, M., Lebenbaum, M., Lam, K., Chong, N., Azimaee, M., Iron, K., & Manuel, D. (2017). Describing the linkages of the citizenship and immigration Canada permanent resident data and vital statistics—Death registry to Ontario's administrative health database. *International Journal of Population Data Science*, *1*(1). https://doi.org/10.23889/ijpds.v1i1.36

Helms, M. M., Ahmadi, M., Jih, W. J., & Ettkin, L. P. (2008). Technologies in support of mass customization strategy: Exploring the linkages between e-Commerce and knowledge management. *Computers in Industry*, *59*(4), 351-363

Jutte, D. P., Roos, L. L., & Brownell, M. D. (2011). Administrative record linkage as a tool for public health research. *Annual Review of Public Health*, *32*(1), 91-108. https://doi.org/10.1146/annurev-publhealth-031210-100700

Moriyama, I. M. (1964). Uses of vital records for epidemiological research. *Journal of Chronic Diseases*, *17*(10), 889-897. https://doi.org/10.1016/0021-9681(64)90160-2

Mule, T. (2021, April 1). Administrative records and the 2020 census. *United States Census Bureau*. https://www.census.gov/newsroom/blogs/random-samplings/2021/04/administrative_recor.html

Newcombe, H. B. (1988). *Handbook of record linkage: Methods for health and statistical studies, administration, and business*. Oxford University Press

PNG tree. (n.d.). *A Broken Chain Free PNG and PSD* [Graphic]. https://pngtree.com/freepng/a-broken-chain_1745989.html

Priestley, M. (2011). *A science of operations: Machines, logic and the invention of programming*. Springer Science & Business Media.

Rosella, L. C., Manuel, D. G., Burchill, C., & Stukel, T. A. (2010). A population-based risk algorithm for the development of diabetes: Development and validation of the diabetes population risk tool (DPoRT). *Journal of Epidemiology & Community Health*, *65*(7), 613-620. https://doi.org/10.1136/jech.2009.102244

Shah, G. H., Lertwachara, K., & Ayanso, A. (2010). Record linkage in healthcare. *International Journal of Healthcare Delivery Reform Initiatives*, *2*(3), 29-47. https://doi.org/10.4018/jhdri.2010070104

Slides Carnival. (n.d.). *Hand-Drawn Timeline Infographics* [Doodles]. https://www.slidescarnival.com/design/hand-drawn-timeline-infographics/129667

Smith, G. (2015). Data linkage: An untapped resource for reducing serious traffic injuries in fast developing countries. *Journal of Local and Global Health Science*, *2015*(2). https://doi.org/10.5339/jlghs.2015.itma.99

Smith, M., & Flack, F. (2021). Data linkage in Australia: The first 50 years. *International Journal of Environmental Research and Public Health*, *18*(21), 1-9. https://doi.org/10.3390/ijerph182111339

Swierenga, R. P. (1990). Historians and the census: The historiography of census research. *The Annals of Iowa*, *50*(6), 650-673. https://doi.org/10.17077/0003-4827.9478