FORS GUIDES
to survey methods
and data management

Data Linkage

Giannina Vaccaro[1] and Elfie Swerts[2]

[1] FORS, LINES, University of Lausanne

[2] FORS

FORS Guide No. 18, Version 1.0

Mai 2022

Abstract:

This guide provides an overview of data linkage, analyses the main advantages, and summarizes the challenges quantitative researchers and data practitioners face when working with multiple data sources. Moreover, this document aims to provide examples of the developments of data linkage in Switzerland as well as inform practitioners about key steps when linking data.

Keywords: record linkage, administrative data, survey records

How to cite:

Vaccaro, G. & Swerts, E. (2022). *Data Linkage*. FORS Guide No. 18, Version 1.0. Lausanne: Swiss Centre of Expertise in the Social Sciences FORS. doi:10.24449/FG-2022-00018

The FORS Guides to survey methods and data management

The FORS Guides offer support to researchers and students in the social sciences who intend to collect data, as well as to teachers at university level who want to teach their students the basics of survey methods and data management. Written by experts from inside and outside of FORS, the FORS Guides are descriptive papers that summarise practical knowledge concerning survey methods and data management. They give a general overview without claiming to be exhaustive. Considering the Swiss context, the FORS Guides can be especially helpful for researchers working in Switzerland or with Swiss data.

Editorial Board

Acknowledgement:

The authors thank Dr. Ursina Kuhn for her review and very useful insights which help improving this FORS Guide and Dr. Pablo Diaz for having reviewed this manuscript and provide us with very useful comments, particularly on the section regarding Legal Provisions relating to Data Linking and Data Protection, and his valuable suggestions to multiples versions of the manuscript. Finally, special thanks to Emilie Morgan De Paula for her valuable comments, and to Dr. Brian Kleiner for his useful tips for improving the document.

Copyright:

1. INTRODUCTION

The growing importance of digitalization of data, better computational power, as well as the emergence of rich data sources, combined with sophisticated statistical methods allow for matching records and bringing together multiple data sources using data linkages. Various stockholders as well as policy makers, clinicians, and researchers demonstrate increasing interest in using linked data from multiple sources to uncover causal effects and answer societal, vital and economical questions. In the social sciences, for example, by linking individuals' records, such as demographic registries over time, one can study topics related to, for example, intergenerational mobility, safety net programs and the impact on longevity issues. In medical studies, the use of linked data helps to measure clinical performance and patient health outcomes to potentially provide better treatment. Performing good and high-quality data linkage analyses is key for advancing research in government, academic institutions, and the private sector.

The pioneering idea of 'data linkage' or 'record linkage' was proposed by Dunn (1946). Newcombe et al. (1959) developed the probabilistic foundations of modern record linkage theory. Later, these were formalized mathematically by Fellegi and Sunter (1969), who showed the possibility of obtaining optimal probabilistic rules when comparing data attributes that were conditionally independent. Since 1990, the development of machine learning techniques, neural algorithms, and artificial intelligence together with the availability of rich-training data have favored accurate estimations of conditional probabilities required in the theoretical foundations. The recent digitalization of records and automatization have reduced or eliminated manual linkage procedures that were prone to error and difficult to reproduce. Computerization has increased the power of data processing, quality checks, consistency of the analysis, and reproducibility of results.

However, linking data sources presents multiple challenges. It requires being aware of specific methodological issues, its technical and data access limitations, especially regarding legal and ethical data protection requirements. Furthermore, legal, and technical challenges are context-specific, and depend on the type of data to be linked (contextual or individual). Access and pre-processing data can be very costly and time-consuming. Ideally, from a methodological perspective, matching records should be: (i) *accurate*, making as few false matches as possible; (ii) *efficient*, creating as many true matches as possible; (iii) *representative*, generating linked samples that resemble the population of interest as close as possible, and (iv) *feasible*, making it possible to be implemented by most scholars given the current limitations of computer power and resources (Abramitzky et al., 2019).

The provision of linked data varies greatly between and within institutions and countries. In some contexts, unique identifiers (IDs) allow one-to-one linkage of multiple datasets.[1] However, in the absence of unique identifiers, two linkage approaches have been developed to reduce the risk of mismatches and probabilistic errors: (a) Deterministic linkages use pre-stablished rules to classify records as belonging to the same or different individuals. (b) Probabilistic linkages assign weights to each pair of records indicating the likelihood of a true match. Both methods have their advantages and limitations, which are discussed later in the document.

[1] For example, OASI Insurance certificate (Switzerland), Social Security Number - SSN (US), Personal Identification Number - CPR (Denmark), Personal Identity Code - HETU (Finland), etc.

The importance of having access to high-quality data enables Switzerland to conduct cutting-edge research at the international level and to inform national policies in the best possible way. This FORS Guide on Data Linkage aims to provide researchers with a general overview of data linkage methods, summarizes the best practices and protocols that should be implemented, and provides some examples of linked data in Switzerland.

2. MAIN ASPECTS OF THE TOPICS

2.1 DEFINITION, ADVANTAGES, AND CHALLENGES OF DATA LINKAGE

Data linkage is defined as the aggregation of data from different sources concerning the same entity or individual. The term *record linkage* has been used since 1959 to "indicate the bringing together of two or more separately recorded pieces of information concerning a particular individual or family" (Newcombe et al., 1959: 954). Linked data (LD) allow the production of new information from existing data, to multiply the value of large data collections, and to answer a very broad range of questions that would not be possible to answer otherwise.

Advantages

Combining different data sources multiplies the value of individual datasets, facilitating the study of large and relevant research questions that were not possible to investigate with single data. For example, a death register has very limited analytical potential, but if it is combined with social or medical information, it can be used to investigate social, or health issues related to life expectancy (Lutz & Swerts, 2020).

By linking multiple datasets one can examine the relationship between different agents in society, improve better identification strategies for causal policy evaluations, and get better knowledge about different subjects. Thus, the richer the dataset (in terms of number of cases as well as of details and extent of information), the more likely data analysts can discover relationships crucial for healthcare, social improvement, and development (Gliklich et al., 2014).

LD makes access to very detailed and rich information easier by combining data from different sources. The availability of large datasets reduces 'information' or 'mis-classification biases' originated by losing information when recoding data; and as a result, improve the quality of data.

The usage of LD allows for obtaining more accurate estimations due to access to more information. Linking registries from different data sources that implement follow-up protocols will also help to keep registers of all variables (Santana, 2009). Accessing diverse sources of information not only increases the sample size because it is possible to include more individuals in the analysis, but also reduces non-response rates by enriching longitudinal information. These increase the probability for later validation of registers, reducing 'selection bias', and generate evidence with a high level of external validity (Harron et al. 2017).

For example, in the field of medicine, by re-entering patients' vital status data recoded in different datasets, we can construct very rich and detailed data. Linking different data sources will not only avoid 'information bias' that would appear if patients that are alive would no longer be under observation (Harron et al. 2017), but will also reduce the 'selection bias' of longitudinal studies by including, for instance, mortality information of the patients initially surveyed (Steck et al., 2015). By linking information from the Swiss censuses from 1990 to 2000 and the Swiss

Childhood Cancer Registry, Steck et al. (2015) geo-coded place of residence at birth and calculated incidence rate ratios (IRRs) to determine the risk of cancer and leukemia in children born relatively near to a central nuclear plant. Also, Spycher et al. (2011) provided a successful example using record linkage in Switzerland.

Finally, the usage of data linkage linked to administrative records is cost-effective: it provides large sample sizes of very detailed information on hard-to-reach population at a relatively low cost. It reduces the burden on respondents by lessening survey dropouts and nonresponse rates, and allows access to very sensitive variables, e.g., on standardized wages and cognitive skills, that would be expensive to access otherwise.

Because of all these advantages, LD allows for a better examination of the relationship of different agents in the society, improves better identification strategies for causal policy evaluations, and obtains better knowledge on different subjects.

Challenges

Despite the multiple advantages of LD, producing and making it available for research and policy evaluation has various challenges.

First, access and pre-processing data can be very costly and time-consuming. Access to data from different sources for linking purposes is limited or sometimes impossible.[2] The lack of clarity and uncertainty related to the access to LD, partially because the need of transparency and non-standard processes from data providers, increase the costs of accessing LD (bureaucratic procedures for signing contracts, etc.). Also, administrative and survey data often contain inconsistent, inaccurate, or incomplete information. Therefore, heavy data cleaning (e.g., blocking and other strategies) and investment in specialized software and high-performance computing capacity that reduce the probability of mismatched identifiers are usually implemented in probabilistic linkage.[3]

Second, privacy concerns and data anonymization constitute a critical challenge for data linkage procedures. Data linkage allows for association of multiple types of information concerning individuals, but at the same time increases the risk of individuals being identified. Well-anonymized datasets can thus be de-anonymized with the addition of new variables which, combined, may allow the identification of individuals. Where this is the case, researchers should be aware that they are processing personal data and that they are therefore obliged to apply the relevant data protection legislation – national, cantonal or even European.[4] In addition, data is usually collected for a specific purpose. Subsequent use for purposes other than those stated at the time of collection is rarely anticipated. This can generate a 'consent bias', commonly known as the "authorization bias or volunteer bias that is described as systematic error in creating treated groups, such that they differ with respect to study outcomes" (Junghans & Jones, 2007). In the survey data linkage process, consent rates vary widely, and might be correlated with observable and unobservable characteristics of surveyed individuals.

Third, guaranteeing privacy and individual anonymity usually involves the implementation of burdensome legal and administrative approval processes. Establishing legal data contracts not only requires researchers to guarantee protecting the information, but also to detail how the

[2] See details in the Federal Act on Data Protection 1992 (FADP).

[3] Blocking strategies are used to increase overall accuracy of the linkage process by restricting the comparison pairs to those likely to match (Harron et al., 2017).

[4] For more information on the applicable legal framework see Diaz (2022).

data are going to be safeguarded.[5] While a data anonymization is required, restricting access to individual identifiers and increasing the administrative and legal burden reduces research potential (Künn, 2015). Practical solutions include accessing LD by remote access, either by publicly available scientific-use files or on-site use for example in highly secured data centers. However, these solutions require more time and resources for researchers to travel to the data centers, and to dedicate exclusively to exploit the data in-situ. This might be usually less comfortable and sometimes more time-consuming because processing the information and performing the analysis in-situ might not be always possible.

Fourth, while accessing to linked data might be costly and difficult for researchers, it might be even harder to grant access to other researchers who are not necessarily related to the project. This can impose serious challenges for analyzing the same linked data for other research purposes and for reproducibility of the results.

2.2 METHODOLOGY

Following the pioneering work of Fellegi & Sunter (1969), we present here the formal concepts behind LD. They established the bases for obtaining *matched* or *unmatched* sets*.*

- *Matched data* will be represented by the union of sets A and B: $A \text{ X } B = \{(a,b); a \in A, b \in B\}$, and
- Unmatched data will be represented by disjoint sets: $M = \{\alpha(a), \beta(b) | a = b\}$, and $NM = \{\alpha(a), \beta(b) | a \neq b\}$.

The first step for linking records of two sets (i.e. A and B), consists in comparing the records, where $\alpha(a)$ refers to the matching variables for entity a in file A, and $\beta(b)$ to the matching variables for entity b in file B. The result of comparing the records will provide indications if the matching variables agree or disagree.

Any record linkage process balances the trade-off between requiring strong evidence of large and good quality data that two records represent the same person and prevents false matches *(Type Error I)* and allowing natural data error between records that represent the same person to avoid missing correct matches *(Type Error II)* (Wiegand & Goerge, 2019). *Type Error I*, also called false positives, occurs when the investigator rejects the null hypothesis that it is actually a good match. *Type Error II* false negatives appear when the investigator fails to reject the null hypothesis when the match is in fact false.

Table 1 summarizes this dilemma:

Table 1: Confusion Matrix - Possible Results for Any Record Comparison

	Records are matched	Records are NOT matched
Person is the same	Successful match	Missed match (Type Error II)
Person is NOT the same	False match (Type Error I)	Successful non-match

Notes. Based on Wiegand and Goerge (2019).

Two methods are commonly used for LD: Deterministic and/or Probabilistic Linkage. Both have their advantages and limitations. While the former is inexpensive in terms of calculation, it is likely to generate mismatches, unless there are unique identifiers. The latter can achieve

[5] As standard when using private data, to guarantee data security researchers usually sign confidentiality agreements that guarantee data safeguard and data destruction after the data analysis.

sufficient linkage quality, but it requires high computational power, and it is prone to errors for administrative data when missing values are present and when using data that vary over time (Harron et al., 2017). Both methods can complement each other, but usually probabilistic linkage is necessary if deterministic matching is not possible.

Deterministic (exact) matching method

In deterministic (exact) matching, the records or matching variables in two (or more) datasets must agree exactly on every character, and therefore 'determines' the result that they correspond to the same entity (Shlomo, 2019). "Deterministic matching assumes there is no error in the way the data is recorded or captured, … and it assumes that all matching variables have equal weights associated to the deterministic matching" (Shlomo, 2019: 48-49). Here, the matching is carried as one-to-one match producing only two potential outcomes: matched or unmatched.

Deterministic matching is generally used when unique identifier numbers are common to the different databases. The greatest strength of deterministic matching consists in its being strictly rule-based. When unique identifiers are not available, other available variables (gender, age, place of residence, etc.) can be concatenated to build a unique identifier number. In this case, rules are articulated and documented so that algorithms can perform minimal calculations and matching routines can be fully automated, reducing overall computation burden.

However, deterministic matching has also limitations. Deterministic matching algorithms are very specific to their underlying datasets, and they usually must be customized for each collected dataset, requiring unique identifiers (such as the Swiss OASI). Moreover, it is necessary to have deep knowledge of data quality and contents for creating a finely tuned deterministic approach that articulates the rules with the elements to be matched (Wiegand & Goerge, 2019).

Probabilistic Data Linkage

Probabilistic linkage attempts to link records of various datasets when there are no unique identifiers, and it is not possible to implement a deterministic linkage. Probabilistic linkage is very effective but at the same time it demands substantial computation efforts to determine the probability that the two sets of identifying variables represent the same unit of population (Oyarzun & Wile; 2016).

A probabilistic data linkage implementation usually involves three different stages (Shlomo, 2019):

- *Pre Linkage:* refers to all the pre-processes involved to edit the data and standardizing matching variables to have the same formats and definitions to be linked. Pre-processing and data cleaning are the most difficult and time-consuming steps in LD, but they are crucial because the success of the matching depends on the quality of the data.

  Shlomo (2019) recommends starting by generating the reference number and adding to each record across the fields to be linked. Then, matching variables will need to be selected following the criteria to be unique, available, accurate, and stable over time. Furthermore, variables involved in the LD must be free of errors. Check particularly spelling variations, missing or miscoded data, standardized formats, and that they have the same characteristics, field length, and coding status across datasets. Identifying duplicate records that refer to the same entity for information integration is key in this process. Bilenko et al. (2003) compare and describe methods (i.e. record linkage, duplicate detection, name matching, etc.) for combining similarity measures for name-matching and identify the key steps for many modern name-matching systems.

To deal with typographical errors, probabilistic data linkage can use phonetic codes and string comparators (i.e. *Soundex* software developed in various statistical packages). Also, string comparator metrics that account for deletions, insertions, and transpositions are useful for data linkage of individuals (Jaro, 1989). Winkler (1990) provides an improved string comparator using weights added to the linkage process.

Another important step in pre-linkage consists in reducing the search space between two datasets by avoiding the comparison of record pairs that are least likely to be matched. This procedure also known as *Blocking Variables*, must be small enough to avoid too many unproductive comparisons, but large enough to prevent records for the same entity failing to link true matches (Shlomo, 2019).

- *Linkage Stage:* In probabilistic linkage, several variables from different data sources are compared and each variable is assigned a weight. This indicates the likelihood or how close the two values of the matching variables are. The sum of the individual variable weights indicates the likelihood of a match between two data sources.

  Following Fellegi & Sunter (1969) and using the terminology employed in the introduction to this section, the aim of the linkage stage is to determine a set of matches (M) and a set of non-matches (NM). The basic *Linkage Rule* (L) can be defined as a mapping from the space Γ, on to a set of random decision functions $D = \{d(\gamma)\}$ where:

  $d(\gamma) = \{P(A_1|\gamma), P(A_2|\gamma), \ldots, P(A_i|\gamma)\}$; $\gamma \in \boldsymbol{\Gamma}$ and $\sum_1^i(\mathrm{P}(A_i|\,\gamma) = 1$

  So, corresponding to each observed value of γ, the *L* assigns probabilities for taking each of the possible actions *(i)*.

  The probability of interest, $P(M|\gamma^j)$, is defined as the marginal probability of a correct match given the record pair j has an agreement pattern γ^j. Following Bayes' theorem this can formally be written as[6]:

  $$P(M|\gamma^j) = \frac{P(\gamma^j|M)P(M)}{P(\gamma^j)} = \frac{1}{1+\frac{P(\gamma^j|NM)(1-P(M))}{P(\gamma^j|M)P(M)}},$$

  where the agreement likelihood ratio is $R(\gamma^j) = \frac{P(\gamma^j|M)}{P(\gamma^j|NM)}$

  The simple procedure consists in maximizing the posterior probability of $P(M|\gamma^j)$ or simply order the likelihood ratio $R(\gamma^j)$ choosing an upper cutoff W^+, and a lower cutoff W^- for determining the correct *matches* and *non-matches*. Then, the optimal *L* given by $F\colon \boldsymbol{\Gamma} \rightarrow \{M, NM, C\}$ maps a record pair j comparison value to a set of three classes: matches (M), non-matches (NM), and minimizes the undecided cases to be clerically reviewed (C), defined as:

[6] The m-probability $m = P(\gamma^j|\,M)$ and the u-probability $u = P(\gamma^j|\,NM)$ are defined as the conditional probability that the record pair j has an agreement pattern γ^j given that is a match (M) and non-matched (NM), respectively.

$$F: \begin{cases} \gamma^j \in M, & \text{if} \quad R(\gamma^j) \geq W^+ \\ \gamma^j \in NM, & \text{if} \quad R(\gamma^j) \leq W^- \\ \gamma^j \in C, & otherwise \end{cases}$$

Furthermore, Fellegi & Sunter (1969) showed that it is possible to estimate the unknown probabilities for each matching and non-matching variable, by decomposing the probability of agreement of record j, with an *Expectation-Maximization Algorithm (EM)* for estimating the parameters and using frequencies of the agreement patterns $(P(\gamma^j|M), P(\gamma^j|NM))$. The *EM* is given by:

$$P(\gamma^j) = P(\gamma^j|M)P(M) + P(\gamma^j|NM)(1 - P(M))$$

where $P(\gamma^j)$ obtains the proportion of the agreement patterns across all possible pairs.

Shlomo (2019) summarizes three key parameters for probabilistic linkage to consider: (i) the quality of the data, represented by $P(M|\gamma^j)$, which indicates the degree to which the information contained for a matching variable is accurate and stable across time; (ii) the change that values of a matching variable will randomly agree; and (iii) the ultimate number of true matches or the marginal probability of a correct match. To ensure the latter exists and therefore a successful LD, high proportion of true matches need to be guaranteed.

- *Post-Linkage:* after implementing the LD process, we need to check the presence of errors Type 1 and Type 2 in the matched and unmatched samples. In other words, there is a need to recall how correctly matched were pairs out of all true matches and specify how correctly not linking non-matches out of all true non-matches (Sensitivity), and then define the number of correctly linked matches out of the total number of linked pairs (Precision) (Shlomo, 2019).

There are multiple modern LD procedures for choosing the best method to be applied in each scenario. Zhu et al. (2015) implemented a simulation study to understand the data characteristics that allow one to determine when to conduct probabilistic or deterministic linkage. In the field of medicine, Gliklich et al. (2014) developed a User's Guide of medical registries with the aim of facilitating the design, implementation, analysis, and interpretation of data registries that help understanding patients' outcomes. In Switzerland, the Institute for Social and Preventive Medicine has developed Medical Registries and Data Linkage (SwissRDL) that creates a registry and collects high-quality data to improve health care.

Multiple software have implemented full packages for Record Linkage:

- In R https://cran.r-project.org/web/packages/RecordLinkage/RecordLinkage.pdf
- In Python https://recordlinkage.readthedocs.io/en/latest/about.html
- In SPSS https://www.spss-tutorials.com/spss-match-files-command/
- In Stata https://journals.sagepub.com/doi/pdf/10.1177/1536867X1501500304

3. PROTOCOL AND BEST PRACTICES FOR LINKING DATA

3.1 LEGAL PROVISIONS RELATING TO DATA LINKING AND DATA PROTECTION

There are no general laws that regulate LD. However, when personal data is (potentially) processed or when the linkage is carried out or subject to the authorization of a federal body, several laws apply, especially regarding data protection. When interested in linking data, three main scenarios must be considered:

1. If *personal data* is contained in the data sets interested to be linked, regulations related to the *Data Protection* apply. These regulations are not specially related to data linkage, but they apply when processing information that can be linked to an identified or identifiable person (art. 3, let. a Federal Act on Data Protection (FADP)). The application of these laws can pose a number of obstacles to research, particularly because they require that individuals be systematically informed of any data collection, including from third parties (art. 18a FADP). Depending on the legal status of the researcher, his field of study, his place of establishment or the geographical location of the data collection, different laws - general or specific - may apply. For example, when the subject of the study is human disease or function, the Human Research Act (HRA) applies. If the principal investigator is employed by a university or a university of applied sciences, cantonal data protection laws apply. For more information see (Diaz, 2022).
2. If the datasets to be linked are anonymous when used alone but their combination allows identification, data protection laws also apply. Indeed, data protection regulations apply from the moment data can be linked to an identified or identifiable person.
3. When LD is carried out by a federal body or is subject to the authorization of a federal body (e.g. the FSO), specific laws apply. Art. 14*a* of the Federal Statistics Act (FStatA, SR 431.01) provides the legal basis for the LD for statistical purposes by federal bodies. Also, the Federal Act of the Harmonization of the Register of Residents and of other Official Registers of Persons (RHA; SR 431.02) explicitly regulates the linkage of data from the Federal Register of Buildings and Dwellings (RBD) and the Business and Enterprise Register (BER). The provision for implementation of the Art. 14*a* of the FStatA is detailed in Legal Bases of Data Linkage of the FSO.

3.2 PROCEDURES TO FOLLOW FOR DATA LINKAGE IN SWITZERLAND

Usually, the linkage of datasets where at least one of the datasets originates from the federal administration for research purposes is authorized in Switzerland by the FSO for research (FStatA). LD is usually carried out by the FSO according to a clear procedure with restrictions, such as limited access to individual IDs to guarantee a very high level of data protection. Usually, if researchers want to link datasets which they (themselves) have collected with data from the federal administration, the researcher must send their own research data to the FSO, who then carries out the linkage: The FSO creates a linkage key (often based on the SSN), links/merges the data, and then makes the anonymized data available.

Under certain conditions, a LD service is provided by the FSO. In this case, the FSO carries out the matches on behalf of third parties (such as federal, cantonal, municipal government, and recognised educational institutions) for non-personal purposes (research, data planning and statistical purposes). These linkages require the signing of a matching and data protection contract.

According to the FSO, the following criteria must be met before implementing data matches on behalf of a third party[7]:

- Goal-oriented: data linkage is only allowed for public statistical or scientific purposes and not for administrative or other purposes. This criterion applies to all clients and requests.
- Legal certainly: the linkage must be implemented only in agreement with legal requirements; and the data included must be in accordance with the Federal Statistics Act (FStatA).[8]

  Data security / data protection must be guaranteed, especially regarding sensitive data. According to the FSO, "only anonymized data that cannot be linked to any individual can be shared. The data must not be de-anonymized or linked to other data. Once the analysis has been made, the data is to be deleted or returned to the FSO" (FSO, 2022: retrieved on April 5th, 2022).
- Methodological requirements: the data to be matched and the resulting dataset should be of sufficient quality, methodologically correct, and suitable for the topic under study.
- Technical feasibility: the data sources to be matched contain identical (pseudonymized) identifiers for the data sources to be linked.
- Implementation: the FSO carries out the matches according to its technical, organizational, and human resources possibilities.

The Swiss FSO has developed a standard procedure for dealing with LD applications, which imply applicants meeting certain requirements (applicants' work for a recognized research institute or for a federal, cantonal or local authority organization, the application is made in a framework of a project that has a statistical (not administrative) objective, the application concerns statistical data of the Federal Administration), and the completion of an application form, available in French or German, and that can be downloaded from FSO website.[9] To carry out data linkage, researchers must apply to the FSO, describing in detail the scope of the project, including the datasets and variables they wish to use.[10] If the data linkage application is accepted, the FSO issues a data linkage and protection contract for the applicant's signature.

In case non-administrative data is involved in the linkage process, it can be done by private practitioners. However, it is important to mention that if *personal data* is involved in the linkage process, all the requirements and legal considerations regarding Data Protection apply (see extensive details in Diaz (2022)).

3.3 SEVERAL INITIATIVES AND CENTRES FOR DATA LINKING

There are many initiatives to facilitate data linking for research purposes and governmental statistical analysis in various fields. For example, in the social sciences, engineering, health,

[7] https://www.bfs.admin.ch/bfs/en/home/services/data-linkages/for-third-parties.html

[8] https://www.fedlex.admin.ch/eli/cc/1993/2080_2080_2080/en

[9] https://www.bfs.admin.ch/bfs/en/home/services/data-linkages/for-third-parties.assetdetail.17084398.html

[10] According to current documentation, the application form should be accurate filled and forwarded to the address verknuepfungen@bfs.admin.ch.

etc. Here, we will present various examples of data linkage initiatives that have been developed for health and social sciences in Switzerland, and around the world.

At the international level, the International Population Data Linkage Network (IPDLN)[11] is a network that connects most of the world's data linkage centres. A list of these centres is available on their website.[12]

Outside Europe, in Australia, the Centre for Health Record Linkage (NSW)[13] assists researchers, planners and policy makers in accessing linked health data on individuals and host a secure, high-performance data linkage system.

In the USA, the Data Linkage Team of the Office of Analysis and Epidemiology of the National Center for Health Statistics[14] has developed a record linkage program to maximise the scientific value of the Centre's population base.

In Europe, multiple institutions have developed various initiatives. Led by the health field, the Norwegian Cancer Registry[15] performs linkages between the cancer registry and other central health registers and other data sources. In the United Kingdom, the HMRC DataLab[16] provides data, including linkage data, and allows authorised researchers to access de-identified HMRC data in a secure government-accredited environment.

In Germany, access to sensitive data and their linkage is ensured by third parties, which are dedicated centres that are accredited by the German Data Forum (RatSWD).[17] There are currently 41 Research Data Centres accredited by RatSWD, which are part of both research institutions and government organizations that host registries and conduct their own research. They provide onsite access to sensitive data for independent academic research. Half of the RDCs also provide data linking.

In France, the National Commission on Computer Technology and Freedom (CNIL)[18] provides secure remote access for researchers to very detailed data. Data linking is facilitated by derogations promulgated by the CNIL and it is based on the involvement of trusted third parties, i.e., institutions that act as third parties between data owners and researchers (as the Secure Access Data Centre - CASD).

Respect to other disciplines, the health sector has advanced data linking efforts in Switzerland and around the world. The Swiss Personalised Health Network (SPHN) Initiative[19], promoted by the Swiss Academy of Sciences (SCNAT). The SPHN aims to create an infrastructure for networking the data of partner institutions, such as hospitals with research centres, universities, etc. Such an infrastructure will enable the interoperability of clinical patient data for research purposes. Health data linkage centres also exist in other countries.

Within the social sciences, the linkhub.ch Initiative[20] regroups several partners from FORS, NCCR on the move, the Centre LIVES, Swiss RDL, Swiss National Cohort and TREE to support the creation of a legal and institutional environment that supports academic and

[11] https://www.ipdln.org/

[12] https://www.ipdln.org/data-linkage-centres

[13] http://www.cherel.org.au/

[14] https://www.cdc.gov/nchs/data-linkage/index.htm

[15] https://www.kreftregisteret.no

[16] https://www.gov.uk/government/organisations/hm-revenue-customs

[17] https://www.konsortswd.de/en/ratswd/

[18] https://www.cnil.fr/en/home

[19] https://sphn.ch/

[20] https://linkhub.ch/

administrative studies based on linked data that combines personal data protection and scientific principles.

3.4 SPECIFIC EXAMPLES OF DATA-LINKAGE AND USAGE IN THE SWISS CONTEXT

A few large data linking projects involving administrative data are underway in Switzerland in fields related to medicine and social sciences. Among these, some well-known initiatives for research access are the Institute for Social and Preventive Medicine (ISPM) and Longitudinal Analysis in the Field of Education data (LABB) projects, which allow for showing the solutions developed for data protection and the contributions of data linking for medical and social issues.

ISPM

The Institute for Social and Preventive Medicine at the University of Bern (ISPM) has developed the Privacy Preserving Probabilistic Record Linkage (P3RL) (Schmidlin et al., 2015), a method for linking personal data without the need for revealing the person identifying information to the linkage centre (trusted third party). The partner organisations for the linkage keep the identifying information, and only encrypted data (a space-efficient probabilistic data structure or Bloom filters) are sent to the linkage centre for the error tolerant probabilistic record linkage. For linkages with P3RL any available identifying information can be used, like person's name, date of birth, date of death or address. The P3RL method consists of three steps:

- *Data preparation:* the files are cleaned and standardised at the sites of the data owner using a pre-processing tool.
- *Encryption:* the trust centre transmits an encryption program (Bloom filters) to the administrators of the partner centres. Data can be imported and encrypted with a key, which is defined by the linkage partners (data owners), without revealing it to the linkage centre. The encrypted data are exported from the tool and sent (without any clinical data for the analyses) to the linkage centre.
- *Record linkage:* the linkage centre performs the error tolerant probabilistic record linkage using record linkage tools. The result of the linkage (link tables with IDs of the linkage partners original data) and a report with additional information about the linkage quality are sent to the partners. The linkage centre then deletes all data of this linkage.

The P3RL method can be used by the community and its principles are well described in the literature. The P3RL tool, which is not freely available, was developed at the ISPM and SwissRDL - a hub for medical registries and data linkage as part of ISPM - which offers the service to perform the P3RL linkage for customers and researchers.

The Longitudinal Analysis in the Field of Education data (LABB)

The Longitudinal Analysis in the Field of Education data (LABB)[21] is a novel and a unique dataset, which provides large register-based longitudinal and harmonized information of persons in Switzerland. Based on both cross-sectional data from official statistics - in education and other domains - and administrative data, LABB compiles relevant information providing rich longitudinal datasets (longitudinal linkage) that allow one to analyse individuals' transitions

[21] https://www.bfs.admin.ch/bfs/fr/home/statistiques/education-science/enquetes/labb.html

through their educational pathways on a large scale, but also in a very detailed way (cross-sectional linkages), taking into account labour market trajectories and life course events (migration status, parental education, educational outcomes, etc.). Variables include education, social origins, migration status, and labour market trajectories, as well as data regarding school transitions.

LABB data and registers are regularly updated and integrated (Babel et al., 2020). Currently, they include data from:

- Education statistics (all levels), from 2012-2018.
- Population and household registers, from 2012-2018.
- Unemployment register, from 2011-2018.
- Social insurance register (CCO/STATENT/SECO), from 2011-2018.
- Structural Enterprise Statistics, from 2011-2018.
- Structural survey, from 2010-2018.

This data is available for academic research and institutional partners under the Data Protection Contract. Access to the data can be requested by contacting directly the Section of Section of Education of the FSO.[22]

4. IMPLICATIONS AND RECOMMENDATIONS FOR PRACTITIONERS

To facilitate the data linkage process, we provide below five suggestions and recommendations:

Recommendation 1 - Take care of data security: Data linking increases the risk of identification of individuals. It is therefore important to respect legal frameworks, and to exercise basic precautions such as keeping data in secure locations. To secure data after being used, usually it is recommended and sometimes required to delete or erase the data after the completion of the project.

Recommendation 2 - Check the presence of variables that allow the matching process. In the absence of a unique identifier, the quality of the database is crucial for matching. Errors in the descriptive variables of the data, names, codes, can greatly reduce the accuracy and quality of the matching. It is therefore important to carefully check the presence of key variables in the data before processing the data linking.

Recommendation 3 - Correctly choose your type of linkage method. The type of method to be used to match the data depends on the configuration of the data. If they have common unique identifiers, then the deterministic method can be used. Without common unique identifiers, a probabilistic method will be used.

Recommendation 4 - Check legal formalities and contracts. Depending on the degree of sensitivity of the data, a contract will have to be signed with the owner of the data to define the modalities of access and the framework of the use and the diffusion of the matched data.

[22] According to the current documentation, for further information regarding LABB data contact directly eduperspectives@bfs.admin.ch.

Recommendation 5 - Facilitate reproducibility of results and finding replications. Document clearly the matching procedure, methods involved, and all processes involved in the data linkage, so that other researchers can replicate your findings.

5. FURTHER READINGS AND USEFUL WEB LINKS

5.1 MAIN SOURCES IN SWITZERLAND

Federal Statistical Office (OFS). Data linkages
https://www.bfs.admin.ch/bfs/de/home/dienstleistungen/datenverknuepfungen.html

Swiss National Science Foundation. Open Science Data
http://www.snf.ch/en/theSNSF/research-policies/open_research_data/Pages/default.aspx

5.2 OTHER CONTACTS IN SWITZERLAND

Swiss Personalized Health Network (SPHN) Initiative https://sphn.ch/

Swiss Data Science Center https://datascience.ch/

Institute of Social and Preventive Medicine at the University of Bern (ISPM)
https://www.ispm.unibe.ch/

Medical Registries and Data Linkage (SwissRDL) http://www.swissrdl.unibe.ch/

The Longitudinal Analysis in the Field of Education data (LABB)
https://www.bfs.admin.ch/bfs/fr/home/statistiques/education-science/enquetes/labb.html

5.3 INTERNATIONAL OFFICES

International Population Data Linkage Network (IPDLN) http://ipdln.org/

Centre for Health Record Linkage (CHRL) of Australia http://www.cherel.org.au/

Data Linkage Team of the Office of Analysis and Epidemiology of the National Center for Health Statistics https://www.cdc.gov/nchs/data-linkage/index.htm

Norwegian Cancer Registry https://www.kreftregisteret.no

German Data Forum https://www.konsortswd.de/en/ratswd/

German Record Linkage Center German Record Linkage Center (uni-due.de)

Scottish Government. Joined-up data for better decisions: Guiding Principles for Data Linkage. https://www.gov.scot/publications/joined-up-data-better-decisions-guiding-principles-data-linkage/

Commission nationale de l'Informatique et des Libertés (CNIL) https://www.cnil.fr/en/home

5.4 MULTIPLE SOFTWARE

In R https://cran.r-project.org/web/packages/RecordLinkage/RecordLinkage.pdf

In Python https://recordlinkage.readthedocs.io/en/latest/about.html

In SPSS https://www.spss-tutorials.com/spss-match-files-command/

In Stata https://journals.sagepub.com/doi/pdf/10.1177/1536867X1501500304

REFERENCES

Abramitzky, R., Boustan, L. P., Eriksson, K., Feigenbaum J. J., & Pérez, S. (2021). Automated Linking of Historical Data. *Journal of Economic Literature,* 59(3), 865-918. doi: 10.1257/jel.20201599

Babel, J., Lagana, F., Falcon, J., Gaillard, L., Strubi, P., & Vasela, J. (2020, November)*.* *LABB (Longitudinal analyses in the field of education) as source to study heterogenerity and inequality in educational trajectories*. Paper presented at Conference organized by the Sociology of Education Research Network of the Swiss Sociological Association (SSA) - SOCEDUC, Bern, CH.

Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., &Fienberg, S. (2003). Adaptive name matching in information integration. *IEEE Intelligent Systems*, *18* (5), 16–23. doi: 10.1109/MIS.2003.1234765

Diaz, P. (2022). Data Protection: legal considerations for research in Switzerland. *FORS Guide* (No. 17). doi: 10.24449/FG-2022-00017

Dunn, H. L. (1946). Record Linkage. *American Journal of Public Health and the Nation's Health, 36* (12), 1412-1416. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1624512/

Federal Act on Data Protection (FADP) of 19 June 1992 (1992). https://www.fedlex.admin.ch/eli/cc/1993/1945_1945_1945/en

Federal Act on the Harmonisation of the Register of Residents and of the Official Register of Persons (Register Harmonisation Act, RHA) of 23 June 2006 (Status as of 1 January 2022) (2006). https://www.fedlex.admin.ch/eli/cc/2006/619/en

Federal Statistical Office (FSO) (2008). *Register of Buildings and Dwellings (RBD): Catalogue of Attributes*. Retrieved May, 9, 2022, from https://dam-api.bfs.admin.ch/hub/api/dam/assets/344172/master

Federal Statistical Office (FSO) (n.d.). *Business and Enterprise Register (BER)*. Retrieved April 4, 2022, from https://www.bfs.admin.ch/bfs/en/home/registers/enterprise-register/business-enterprise-register.html?msclkid=6d6de68fcf9311ecb1ab2cd42afb82e1

Federal Statistical Office (FSO) (2022). *Data linkages for third parties*. Retrieved April, 25, 2022, from https://www.bfs.admin.ch/bfs/en/home/services/data-linkages/for-third-parties.html

Federal Statistics Act (FStatA) of 9 October 1992 (Status as of 1 January 2016) (1992). https://www.fedlex.admin.ch/eli/cc/1993/2080_2080_2080/en

Fellegi, I. & Sunter, A. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328), 1183-1210. doi: 10.1080/01621459.1969.10501049

Gliklich, R. E., Dreyer, N. A., & Leavy, M. B. (2014). Pregnancy Registries. In R.E. Gliklich, N.A. Dreyer, & M.B. Leavy (Eds.), *Registries for Evaluating Patient Outcomes: A User's Guide [Internet]* (3rd ed.). Agency for Healthcare Research and Quality (US). https://pubmed.ncbi.nlm.nih.gov/24945055/

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., & Goldstein, H. (2017). Challenges in administrative data linkage for research. *Big Data & Society*, 1-7. doi:10.1177/2053951717745678

Jaro, M. (1989). Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, *84* (406), 414-420. https://doi.org/10.2307/2289924

Junghans, C. & Jones, M. (2007). Consent bias in research: how to avoid it. Heart, 93 (9), 1024-1025. doi: 10.1136/hrt.2007.120113

Künn, S. (2015). The challenges of linking survey and administrative data. *IZA World Labor: Evidence-based policy making*, 1-10. doi: 10.15185/izawol.214

Loi fédérale sur la protection des données (nFADP) du 25 septembre 2020. https://www.fedlex.admin.ch/eli/fga/2020/1998/fr

Lutz G., & Swerts, E. (2020). *Accessing and linking data for research in Switzerland*, FORS-Centre de compétences suisse en sciences sociales. Retrieved May, 10, 2022, from https://api.swiss-academies.ch/site/assets/files/23916/report-data-access-and-linking-11-2020-final-1_02-1.pdf

Murat, S., & Brog. A. (2022). *Package 'RecordLinkage'. In: Record Linkage Functions for Linking and Deduplicating Data Sets (R package version 0.4-12.3)*. The CRAN-R package repository, from https://cran.r-project.org/web/packages/RecordLinkage/RecordLinkage.pdf

Newcombe, H. B., Kennedy, J. M., Axford, S. J., & James, A. P. (1959). Automatic linkage of vital records Computers can be used to extract "follow-up" statistics of families from files of routine records. *Science*, *130* (3381), 954-959. http://www.jstor.org/stable/1756667

Oyarzun, J., & Wile L., (2016). *Aperçu du couplage d'enregistrements de données d'entreprises à Statistique Canada: Comment coupler les enregistrements «non coupables»*. Statistique Canada. https://www.statcan.gc.ca/fr/conferences/symposium2016/programme/14741-fra.pdf

Santana, P., Santos, R., & Nogueira, H. (2009). The link between local environment and obesity: a multilevel analysis in the Lisbon Metropolitan Area, Portugal. *Social science & Medicine*, *68* (4), 601-609. doi: 10.1016/j.socscimed.2008.11.033

Schmidlin, K., Clough-Gorr, K. M., & Spoerri, A. (2015). Privacy preserving probabilistic record linkage (P3RL): a novel method for linking existing health-related data and maintaining participant confidentiality. *BMC Medical Research Methodology, 15* (46). doi: 10.1186/s12874-015-0038-6

Shlomo, N. (2019). Overview of Data Linkage Methods for Policy Design and Evaluation. In N. Crato, & P. Paruolo (Eds.), *Data-Driven Policy Impact Evaluation* (pp. 47-65). Cham, CH: Springer. https://doi.org/10.1007/978-3-319-78461-8_4

Spycher, B. D., Feller, M., Zwahlen, M., Roosli, M., von der Weid, N. X., Hengartner, H., Egger, M., & Kuehni, C. E. (2011). Childhood cancer and nuclear power plants in Switzerland: a census-based cohort study. *International Journal of Epidemiology*, *40* (5), 1247-1260. doi:10.1093/ije/dyr115

Steck, N., Spoerri, A., & Egger, M. (2015). Appariement et protection des données de santé: une contradiction?. *Bulletin des médecins suisses – Schweizerische ärztezeitung – Bollettino dei medici svizzeri*, *96* (5051), 1837-1840. https://doi.org/10.4414/bms.2015.04207

Wasi, N., & Flaaen, A. (2015). Record linkage using Stata: Preprocessing, linking, and reviewing utilities. *The Stata Journal, 15* (3), 672-797. https://journals.sagepub.com/doi/pdf/10.1177/1536867X1501500304

Wiegand, E. & Goerge, R. (2019). *Record linkage innovations for the human services. The Consortium stimulating self-sufficiency & stability scholarship*. University of Chicago. https://www.chapinhall.org/wp-content/uploads/PDF/Record-Linkage-Innovations-for-the-Human-Services.pdf

Winkler, W. (1990). *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.* US Bureau of the Census. Stat. Research Div., Washington. https://files.eric.ed.gov/fulltext/ED325505.pdf

Zhu, Y., Matsuyama, Y., Ohashi, Y., & Setoguchi, S. (2015). When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. *Journal of Biomedical Informatics*, *56*, 80-86. doi : 10.1016/j.jbi.2015.05.012