Accessing and linking data for research in Switzerland

November 2020

This report has been commissioned to FORS and linkhub.ch by the Swiss Academy for Humanities and Social Sciences (SAGW) and Swiss Academies of arts and sciences. linkhub.ch is a joint initiative from different research projects and institutions to facilitate access to private and public data for research.

This report addresses the importance, current practices, and the legal basis for access and linkage of administrative and sensitive data (in particular, personal data) for research in Switzerland. It provides a basis for the development of a research-friendly institutional and regulatory framework that will not only enable wider and more comprehensive access to data for research but also facilitate data linking while respecting the privacy and data protection rights of individuals.

It is intended for the competent bodies of the education, research, and innovation (ERI) domain (SERI, SNSF, swissuniversities, Swiss Academies) and other stakeholders in order to provide a basis for discussion on how to improve the conditions for high-quality research in Switzerland. The report is also intended for the Federal Offices concerned with the subject, as well as for scientific communities that conduct research on the basis of personal data, such as the social sciences and human medicine.

The primary authors are Elfie Swerts and Georg Lutz from FORS; the legal parts are based on an analysis by Anna Kuhn and Nora Zinsli from SWITCH. This report was then consolidated with advisory board that consisted of Ben Jann, Kurt Schmidheiny, Adrian Spörri, and Markus Zürcher. Various other colleagues from the SNSF, swissuniversities, the Swiss Academies and different Swiss Federal Offices have provided valuable inputs over the last years on different aspects of this report.

Executive Summary

In the last decade, the empirical basis of research has changed fundamentally. For a long time, research was based on data collected specifically to gain scientific knowledge. In recent years, more and more data is being used for research: data which is generated by the digitization of many areas of work and life.

Access to high-quality data has always been a key strategic factor in advancing high-quality research within the administration, academic institutions, and the private sector. For data that does not include personal information, access is not problematic. However, there are many data collections in the public and private sectors that contain personal data. While such data collections are often interesting on their own, they become even more valuable and informative when individual characteristics can be linked. Linked data from different sources increases accuracy, helps save on costs, and reduces the burden on respondents.

However, public and private data are often not FAIR (findable, accessible, interoperable, and reusable). Comprehensive metadata and documentation do not exist or are not publicly accessible, access to data is complicated or denied altogether, and if data is accessible at all, the use of the data and the linking of the data can only be done to a very limited extent. In the end, researchers often have to destroy the data again, which violates the principle of reproducibility of scientific results. These conditions urgently need to be improved.

The contradictory public discourse is also of little help. On the one hand, there is a political demand for better access, especially to administrative data, and on the other hand, there is an increased demand for more data protection. Discussion on these two topics is rarely conducted simultaneously. However, as shown in connection with the COVID-19 pandemic, the discussion should not be about whether administrative as well as private data should be made accessible for research, but rather about how best to do this.

With a favorable institutional and legal framework, it would be possible to improve access and data protection simultaneously. Institutions and processes need to be designed in such a way that the task of data linking is disconnected as much as possible from the access and analysis of the linked data. Access to highly sensitive data needs to happen in a secure and restricted environment. This report lays out a concrete first proposal on how the institutional and legal framework could be designed to improve data access and data linking.

Follow-up activities are needed. Academic institutions and policy makers must recognize the need for a joint strategy, and the political authorities should develop and implement an institutional and legal framework favorable for researchers to access and link data. This is essential to provide conditions in which Swiss research can remain competitive at the international level in the future.

Table of contents

1 Introduction: Changing empirical foundations of research

Access to high-quality data for research is an important strategic resource and a key factor in advancing research within the administration, academic institutions, and the private sector. In the last decade, there has been an important shift in the empirical foundation of research. For a long time, research had been based on data collected specifically to produce scientific knowledge. Increasingly, researchers rely on data that is produced through the digitalization of many parts of our lives. Digitalization has led to a massive increase in data. Many administrative and production processes have become digital. Every interaction with the internet creates multiple data points. With the internet of things, a next wave of data multiplication is on the horizon and will create further digital information on many everyday activities. Technological improvements in data collection processes and more elaborate research methods and designs also allow for new and more advanced data analysis.

Much of this data is personal data: digital information on identified individuals. While such data collections are often interesting in themselves, they become even more valuable and informative if individual characteristics can be linked: linked data from different sources increases the value for knowledge advancement, accuracy, and cost-effectiveness while reducing the burden on respondents. We can see in current times how policy-making depends on access to good data. Since the beginning of the COVID-19 pandemic, there has been a massive need for information from the public and private sectors to monitor or analyze many aspects of the current situation. Moreover, this information is crucial because far-reaching measures have to be taken on short notice in times of crisis. In contrast, governments as well as independent researchers struggle to obtain access to data, and there are many practical and legal issues related to such access. The current situation has shown, in a stark way, the shortcomings of the current legal and institutional framework.

The public discourse on data protection and data access is currently disconnected. On the one hand, there is a political demand for better access, especially to administrative data. Open government data has been an important principle for some time, and it has become widely acknowledged that personal data and data on companies should also be shared and, to some extent, linked in order to reduce the burden on those who have to provide the data. Data should only be collected once and not multiple times by different governmental bodies. On the other hand, there is an increased demand for more data protection. Using and linking existing data can indeed further increase the sensitivity of personal data because of the increased harm risked by the disclosure of more complete and combined sensitive individual information.

This current political discourse has led to unsatisfactory outcomes. While the value of high-quality data is undisputed in principle, the linkage, access, and archiving of linked data faces many challenges. Administrative as well as private data are far from being FAIR (findable, accessible, interoperable, and reusable) for researchers. Comprehensive metadata and documentation do not exist or are not publicly available, access is complicated or denied, and, if data is accessible, the use of the data and the linking of the data can only be done within a very restricted framework. In the end, researchers often have to destroy the data again, which violates the principle of reproducibility of scientific results.

Access and linking of public or private data are currently involved in a struggle to shift the boundaries between access and protection. This is not a fruitful discussion. As a starting point, admitting that access to such data for research purposes should be granted in principle, the discussion should evolve toward how to integrate both the need for data access and linking for research on the one hand, and the need for better security and data protection on the other. *Access* in this context does not mean

that all researchers can simply download all data. This is neither realistic nor desirable, and downloading raw data should, in many cases, be prohibited because data cannot be protected in this manner. Instead, it means that data is, in principle, usable for research purposes with all the necessary data protection safeguards in place. The way to go is not to try to limit data access and linking as much as possible but to reflect on better ways to provide data security and data protection for administrative and private as well as linked data. It is essential for Switzerland to overcome this dichotomy with regard to the importance of the access to and linking of data in the future.

Better data access and improved linking possibilities are beneficial in many ways. Generally, it allows studying relevant political, societal and economic questions with higher precision and therefore producing important knowledge for policy- and decision-making. It is also important for the Swiss research environment more generally, because access to high-quality data is essential for the competitiveness of the Swiss research area. Clarity on how data can be accessed and used is relevant to plan and fund research which is in the interest not only of researchers but research funding organization. In addition, a more systematic exchange with data producers from the private and public sector with researchers allows mutual learning and knowledge creation on how to handle, protect or analyse data, and helps to improve data literacy and data quality over time.

Switzerland should therefore develop a comprehensive legal, institutional, and technical environment to respond to the need for data access, processing, and storage, while ensuring the protection of individuals' privacy, to guarantee its independence with respect to its data, as well as the trustworthiness of the data, in such a way that:

- data remain under the control of the data owner (e.g., administration, research institutions, researchers, and companies);
- data protection is ensured to the highest degree possible and sensitive data is not stored on private infrastructures (e.g., companies), publishing houses, or journal websites; and that
- privacy and data rights and intellectual property rights are protected within Switzerland. Private infrastructures located outside the country follow different protection rules, which is especially problematic for sensitive and linked data.

This report thus aims to provide the basis for the development of a research-friendly institutional and regulatory framework that allows broader and more comprehensive access to data for research and a system for data linking and sharing that respects the privacy and data protection rights of individuals. It examines the potential offered by data for research in Switzerland and the difficulty researchers experience in accessing it by:

- *(i)* highlighting specificities of administrative and private data as well as challenges to accessing data in a broad sense (Chapter 2);
- *(ii)* examining the current legal framework governing access to data and data linking for research and the current practices of researchers, administrations, and private companies regarding data access and linking in Switzerland (Chapter 3);
- *(iii)* presenting practices and the legal environment of accessing and linking data in selected countries (Chapter 4); and
- *(iv)* proposing solutions to improve the Swiss institutional and legal framework for accessing and linking data (Chapter 5).

2 Specificities of administrative and private data and the challenges of using such data

When researchers design their data collection to answer a specific question, they have a great deal of control over the data collection process, which allows for assessing the data quality from beginning to end. In addition to such data produced by researchers, non-research data has become increasingly important because of its advantages for the research process:

- A large quantity of data from the administration is available that has a *high level of detail and high accuracy on an entire population.* This allows researchers to address new and important questions with high precision. For example, detailed data on income from social security or data from tax declarations enable research on the economic situation of individuals in a way that is almost impossible or would be very costly to do through designed data. Many people would not be able or willing to share such information in detail nor to go back in time and provide a history. In addition, working with full populations allows researchers to avoid many of the potential sources of bias in the research process when working with samples related to non-response and selective inclusion of individuals.
- *Data from different sources can be combined and linked,* which multiplies the value of large data collections. A single data collection often has very limited information, but if different sources can be combined, very rich information can be generated. For example, a death register in itself has very limited analytical potential. However, if a death register can be combined with medical or social information of individuals, it can facilitate the study of a large number of relevant research questions, e.g., what social or health factors are related to life expectancy. In this respect, accessing private data or linking administrative or research data with data from companies would be very desirable too.
- New data *allows researchers to study new research questions*. A great deal of data exists that has not been available before, and supply often creates demand and stimulates new research ideas. An entire new profession—that of data science -- has been created, which focuses on finding answers to questions that have not even been asked at the beginning of the research process.
- Using existing data *decreases the burden on respondents*. Data that has been collected once does not need to be collected directly from individuals or companies again if data sharing is allowed. To decrease the burden is also a request from the political realm. The “once-only” principle demands individuals and companies not be asked multiple times to provide the same information but rather that information provided once be shared and reused by others.
- Designed data can be very costly, while the *use of existing data can lower costs*. Using existing data is not cost-free, because a substantial amount of time has to be dedicated to making the data usable for research. This includes data cleaning, data processing and data harmonizing, because the data is not necessarily in the format needed for analysis. Nevertheless, however often these steps may be required, it is still less costly than collecting similar data from scratch.
- Usage of existing data and a systematic and open exchange between data producers and data users *allows for improvement of data quality and data literacy* more generally. This is also important for data producers because it enhances the value of their data and helps those who collect and process data to reflect on quality. Producing high-quality data and handling data is a learning process. The value of data and potential problems are often only detected once data is used in different ways. Data is never perfect. Some data producers may therefore see a reputation risk to

make their data available. However, over time it should become clear that in the long run sharing data will be beneficial because it helps improving data collection and processing.

However, using such data for research entails different practical challenges:

- Research, particularly in the social and medical sciences, is often based on *individual data that can be sensitive*. Compared to designed research data that is often based on samples, data from existing sources frequently includes an entire population and, as such, results in a higher potential risk and makes it easier to identify individuals. As a consequence, data protection concerns increase substantially. Compared to samples, data on entire populations is also more prone to misuse for commercial purposes.
- Administrative and private data are usually *not designed to answer specific research questions*; instead, the data are collected and stored for other reasons. In order to use them for research, researchers have to clarify how representative or biased the data are for a given population, what conclusions can be drawn, and what the limitations are. This is often difficult since the details of how the process of data collection happened remain unknown.
- In this regard, in some cases, *processing costs can be very high* to make such data usable for research. Since, at some point, untrained individuals have entered data, there might be many mistakes, or systems of data entry that are imperfect or have changed over time may also produce errors. Researchers often lack detailed knowledge of how data was entered and processed, which makes it challenging to use administrative data.

In addition, administrative and private data are far from being FAIR, and accessing and linking sensitive data remains complex for researchers in many ways:

- *Structured metadata and documentation may not be publicly available or may be lacking; hence, data is not easily findable.* Some data producers often do not release publicly or in great detail information on the data collections that they hold, especially in the private sector but also with respect to administrative data. Even if this is known, structured and rich metadata or comprehensive documentation of administrative or private data often does not exist or is not publicly available. Furthermore, if documentation does exist, it does not necessarily include all the relevant information a researcher would need to use such data.
- *Access and linking to data from different sources are limited* or, sometimes, impossible for a multitude of reasons. In some cases, accessing and linking data is not legally possible. Especially with public data collections, these can only be carried out with a clear legal basis and for a specific purpose. This purpose often does not include using data for research. In addition, data owners (whether from the governmental, research, or private sector) are often unwilling to transmit their data to another institution for security reasons, even if that institution is trustworthy. Data owners are afraid of data protection breaches that will then backfire on them. Or, data owners may be worried that their data will be misinterpreted or that someone might find errors and flaws in a data collection. Private companies may also be unwilling to make data available or even disclose the data they possess because data is a strategic resource for a company, and some data may be collected and stored without citizens' knowledge and, in some cases, also sold to other private companies for commercial purposes.
- *Procedures for accessing and linking sensitive data can be lengthy, complex, and uncertain*, slowing down or even interrupting ongoing research projects. The slowness and partial lack of clarity regarding the procedures to access data (which sometimes takes several years) means that both

researchers and data providers have to invest much in resources in the process. The uncertainty around whether, when, and how access to data is given is also problematic when planning a research project. Funding agencies routinely request a guarantee that data access should be possible, while data owners often may not be able or willing to give such a guarantee. They might only look into doing so once funding is available, or might also reconsider their consent to grant access later on in the process.

- *There is a lack of standards and secure steps and procedures in the chain of sharing, linking, and processing highly sensitive data* that can guarantee data security. This prevents, in some cases, data owners from making data available.
- More specifically, for *private data, companies are unwilling and also have no obligation or incentive to share their data*. Due to transparency and open-government policies, the administration is more open to the idea of data sharing than private companies in many countries. Furthermore, the steps for accessing and linking private data are not harmonized but vary from company to company, and there is no dedicated contact or referent institutions to whom to turn to for guidance.
- *The reuse of data for research work is often very restricted* mainly because of data protection issues, and data often has even to be destroyed at the end of the research process. This is problematic in many ways. It is an important scientific principle that results should be reproducible. If the underlying data sources are not accessible or even destroyed, this is hard to fulfill. In addition, many projects invest significant resources in making data useable for research. These resources are wasted if it is not possible to reuse existing datasets or even expand on existing data.

There are many issues related to administrative or private data in general; however, there are also some aspects specific to linking personal data for research.

- Foremost, linked data increase the privacy concerns substantially. On the one hand, there is more information on an individual available and often, as a consequence, also more sensitive information. When linking data from different sources, the risk of data privacy is even greater because all data providers have to send their data (including identifying and descriptive variables) to a unique institution. In order to link data, personal identifiers need to be available in the different datasets; so, by definition, individuals are identifiable. This can be a unique number (such as a social security number) or personal information, such as a name, address, birthday, place of birth, etc., that allows for probabilistic matching. Obviously, much of the data produced by the administration and private companies are personal and sometimes sensitive data. Moreover, although new and efficient methodologies are being developed, big data, whether produced by administrations or companies, are difficult to anonymize.
- Linking data may *be complicated and incomplete*. The linking process can be flawed or difficult. Datasets to be linked, even when they theoretically cover the same population, may be incomplete or have errors. Moreover, identifying variables may not allow for a complete match or lead to unclear results, especially when probabilistic matching is performed. This means that combining two datasets can create many cases where merging and matching information is not possible. Researchers need to understand this linking process in detail and the bias in the data that may be produced.
- *Data from different sources may not be readily compatible* because they use different formats, codes, and concepts. This means that researchers have to check the content of the datasets carefully and might need to invest substantial effort in data preparation.

- The *storage of data for reuse* for secondary analyses or for replication may be limited. This violates the principle of reproducibility of scientific results significantly, which is a fundamental principle that is also increasingly enforced by funders or journals. Since the linked data cannot be stored, the data must be retransmitted and relinked for reuse, preventing subsequent analyses from being replicated under the same conditions as the first analysis.

3 The legal and practical situation in Switzerland regarding accessing and linking data[1]

This chapter, which outlines the different aspects of the legal framework in more detail, is relevant to administrative data. We focus on administrative data because it has a different status from private data with respect to the legal framework. Indeed, administrative data collections need a legal basis, as a State cannot collect data without a legal basis. On the other hand, private data collections do not need a specific legal basis, although they are subject to some regulations, mainly the Federal Data Protection Act.

In addition to the Federal Data Protection Act, which regulates the processing of personal data for federal authorities and private entities, three other major acts regulate the procedures related to data, from their collection to their use. The Federal Act on Freedom of Information in the Administration (FoIA/*Öffentlichkeitsgesetz*) is also relevant for government data. It regulates the rights of access to official documents produced by the administration. The Federal Statistics Act (FStatA) provides the federal administration with the statistical principles it needs to fulfill its functions and regulates data linkage. Finally, specific laws exist for research, such as the Federal Act for the Promotion of Research and Innovation and the Act on Human Research.

We detail the content of each of these acts (3.1) before focusing on the procedures for researchers to access and link administrative data (3.2).

3.1 The legal framework in Switzerland

The following laws are relevant for data access and data linking for research.

The Federal Act on Freedom of Information in the Administration (FoIA/Öffentlichkeitsgesetz)

The Federal Act on Freedom of Information in the Administration (FoIA) provides any person the right to access documents[2] (information) of the federal authorities and to request information on the contents of documents (FoIA, art. 6). Natural persons may request copies of documents. The right of access to official documents can be restricted, postponed, or denied if this is contrary to overriding public or private interests, as listed in article 7, paragraph 1 of the FoIA; there is no entitlement to access, for example, if the internal or external security of Switzerland could be endangered by its guarantee or if

[1] The legal analysis is largely based on an outline that FORS commissioned from the legal services of SWITCH (Anna Kuhn and Nora Zinsli).

[2] An official document is, basically, any information (a) that has been recorded, regardless of the medium, (b) that has been retained by the authority that issued the same or to which it has been communicated, and (c) that concerns the execution of a public function. Documents that are used by an authority in a commercial capacity, have not been issued, or are intended for personal use are not deemed to be official documents (FoIA, art. 5).

the access is likely to reveal professional, business, or manufacturing secrets. Access may also be restricted, postponed, or denied if its granting would impair the privacy of third parties; in such a case, however, the public interest in access may prevail (FoIA, art. 7, para. 2).

The procedure for access to official documents is governed by article 10 of the FoIA. A request for access to official documents must be addressed to the authority that issued the document. If the official documents in question contain personal data, the authority shall consult the subject of the data, who is granted ten days to submit comments (FoIA, art. 11). Within 20 days at the latest, the authority shall give its opinion on the application for access (FoIA, art. 12). If access to the official documents is refused or restricted, there is the possibility of submitting a request for conciliation (FoIA, art. 13). The Federal Data Protection and Information Commissioner bears the duty, among others, of conducting mediation proceedings (art. 13) and making recommendations (FoIA, art. 14) in the case of an unsuccessful mediation.

The FoIA does not constitute a legal basis for an authority to grant access or for a research institution to request access to data. The FoIA is designed to respond to individual case requests. Nevertheless, the FoIA is the only law that regulates general access to data. For this reason, the FoIA appears to be the law that, through amendments, is suitable for regulating and thus enabling access to data for research institutions. The circumstances under which such access to data should be facilitated and the challenges to be considered are set out in the following subsections.

The Federal Statistics Act (FStatA)

The FStatA's purpose is to provide the federal administration with the statistical principles that it requires to fulfill its duties; publish statistical results for the cantons, the municipalities, the economy, the private sector, representatives from civil society, and the general public; promote national and international cooperation; and ensure data protection within federal statistical bodies, among others (FStatA, art. 1).

Article 14 of the FStatA restricts the use of statistical data for any other purpose unless this is stipulated in another federal law or with the express consent of the person concerned. Therefore, the FStatA cannot serve as a legal basis for the processing of data carried out by research institutes.

Interestingly, article 14a of the FStatA regulates data linking, which includes that research institutes can potentially also process data. Article 14a of the FStatA stipulates that data needs to be rendered anonymous if linked. In the event that data linking involves data considered especially sensitive or that data links generate personal profiles, the linked data must be deleted upon completion of the statistical analysis. Therefore, article 14a of the FStatA can be considered a threshold for using linked data in the future that may be the subject of forthcoming discussions.

Federal Act on Data Protection (FADP current version)

The Federal Act on Data Protection (FADP) regulates the processing of personal data for federal authorities and private entities. Article 4 of the FADP sets out the general principles for processing personal data, which are, among others, that personal data may only be processed lawfully and that such processing must be carried out in good faith and be proportional. Personal data may only be processed for the purpose indicated at the time of collection, that is evident from the circumstances, or that is provided for by law.

Federal authorities, universities, or public institutions may process personal data for research purposes, in accordance with article 22 of the FADP. Therefore, personal data can only be processed for

research purposes if anonymized (as soon as the purpose of the processing permits), data can only be transferred with the consent of the authorities, and research results can only be published in an anonymized way so that single individuals cannot be identified. Furthermore, according to the Federal Data Protection and Information Commissioner, a legal basis, i.e., in federal law, is needed for the federal authorities and public bodies subject to the FADP to process personal data for research reasons, which is based on article 17 of the FADP.[3]

Article 13, al. 2, lit. e of the FADP also regulates the processing of personal data for institutions other than public bodies, such as federal authorities, universities, or public institutions, to which this article is not applicable. It states that for the purpose of research, personal data may be processed in such a manner that the data subjects may not be identified when the results are published (FADP, art. 13, para. 2, lit. e).

With these regulations, a legal basis is provided for processing personal data for research purposes (as described under certain conditions). The FADP does not constitute, however, a legal basis for access to data for research purposes.

Revised Federal Act on Data Protection (E-FADP, ongoing revision process)

The revised E-FADP includes article 27, paragraph 4, lit. e, which is similar to the current article 13, paragraph 2, lit. e in the FADP. The processing of personal data for research purposes is allowed if this data is anonymized, and results are only published in a way preventing any link to an individual. The scope of article 27 of the E-FADP is similarly restricted to private entities and excludes public bodies.

Article 35 of the E-FADP further stipulates that the processing of personal data for non-personal purposes is possible if sensitive personal data is anonymized, the processing is covered by the original purpose, or the results are anonymized (equivalent to FADP, art. 22).

Therefore, the E-FADP does not constitute a new basis for access to the data of the federal administration for research purposes. The provision according to which personal data may be processed for research purposes under certain circumstances does not go so far that a right of access to data can be derived from it.

Federal Act on the Promotion of Research and Innovation (RIPA)

The Federal Act on the Promotion of Research and Innovation applies to research bodies that use federal funding for research and innovation. Consequently, this act does not apply to privately sponsored research or institutions governed by cantonal law. The act defines its purpose as aiming to encourage scientific research, encourage science-based innovation, support the analysis and exploitation of research results, ensure cooperation between research bodies, and ensure the economical and effective use of federal funding for scientific research and science-based innovation (RIPA, art. 1). The RIPA, however, does not constitute a legal basis for research institutes to access data for research purposes.

3.2 The current practices of data access and data linking for administrative data

Federal as well as cantonal administrative units already make anonymized personal data available for research in many cases. While different departments grant access occasionally to their data sources, the Federal Statistical Office (FSO) has guidelines and procedures in place for access to personal data

[3] See https://www.edoeb.admin.ch/edoeb/de/home/datenschutz/statistik--register-und-forschung/forschung/datenschutz-und-forschung-im-allgemeinen.html (visited 08.25.2020).

("*Einzeldaten*"). Researchers have to sign a data contract that limits the usage of data to a maximum of five years, and data has to be destroyed after the end of a project. Researchers also have to make sure that they respect data protection. The FSO also provides metadata on the databases it produces and has a federal mandate, issued by the Confederation, as part of its policy of the "once only" principle, to pursue the standardization and harmonization of federal administration data in collaboration with the other departments. This implies the construction of a data catalogue by the FSO with information on data stored throughout the entire federal administration. In spite of these existing catalogues and initiatives, many challenges remain with regard to ensuring that metadata is comprehensive, widely usable, and interoperable.

Federal units also make non-anonymized personal data available. For example, the FSO provides address samples from the "*Stichprobenrahmen für Haushalts- und Personenerhebungen*" for specific research projects. Data distributed only contains limited information, such as names, addresses, birthdays, or marital status, with little potential to harm posed to individuals. Researchers have to sign a data contract, and they are only allowed to use the data for research projects; they are not allowed to link data, and the data needs to be destroyed at the end of any project.

The FSO has also enabled data linking not only for its own use but also for many research projects based on article 14 of the Federal Statistical Act. Data linking has its own regulation ("*Verordnung des EDI über die Verknüpfung statistischer Daten*") and guidelines and is conducted following a clear procedure and with restrictions, such as limited access, in order to ensure a very high level of data protection. All requests have to be approved by the director.

Data that can be linked through this procedure include mainly federal data, and, usually, data stored at the FSO is included. Data files to be linked from different sources have to be sent to the FSO, which then creates a new linking key (often based on the social security number), links/merges the data, and then makes the anonymized data available. External collaborators can also be involved; however, they need to be working inside a secured room at the FSO where up- or downloading data is not possible.

While many different datasets for many projects have already been linked in the past,[4] the current framework also has some shortcomings for researchers:

- The administrative procedure to access data is still burdensome. Researchers have to describe in detail which datasets and variables they want to use, and they have to specify variable names. This is sometimes difficult since the existing datasets within the federal administration are not all very well described, especially for data that is not collected by the FSO.
- All data owners have to sign a contractual agreement, which can take a great deal of time if several data owners are involved, and data owners can also decide not to make their data available individually. This is the case in some instances (e.g., health data and other data). This uncertainty makes it difficult to plan a research project and obtain funding.
- Data linking is only possible if at least one federal dataset is involved. If this is not the case, linking of data is not possible or not regulated at the federal level.
- If researchers want to link data that they have collected with data from the federal administration, they have to send their own data to the FSO. This can be problematic if this includes highly sensitive data, such as information on religion, ideological or political views, political or trade union activities and memberships, race, health, intimate information, social welfare, or crimes.

[4] https://www.bfs.admin.ch/bfs/de/home/dienstleistungen/datenverknuepfungen/allgemein.html

While in this chapter, we have only looked at administrative data, much less is known or defined about the usage of private data for linking. Some companies, such as Google, Facebook, and Amazon, collect and combine a large amount of data from different sources, also from many Swiss citizens. Some of these data collections are of great public interest since they contain interesting and relevant individual information, but the access to such data is not regulated and very difficult in practice. No standards or even principles governing how such data should be used for research exist, and despite of a great public interest, access to data depends entirely on the willingness of the private data owners to make data available.

3.3 Ongoing developments

Several initiatives have recently been launched in Switzerland to develop data access and linkage for research, which demonstrates their importance and added value for society. For research data, the State Secretariat for Education, Research, and Innovation (SERI) has commissioned swissuniversities to develop a strategy for open research data. This will be completed in 2021.

For health data, the Federal Office of Public Health (FOPH) is currently preparing a report for the "Postulate 15.4025 Ruth Humbel" for the better use, sharing, and recording of health data. For administrative data, as mentioned in section 4.1, the Confederation has launched a "once only" principle policy, which aims to enable the wide reuse of the datasets produced by the administration ("*Nationale Dateninfrastruktur*"). As part of this policy, the FSO is implementing a national data management program based on an interoperability platform that hosts a data catalogue providing information on all the data produced by the federal and cantonal administrations. This policy has also resulted in the construction by the FSO of a center of competence in data science (*Kompetenzzentrum für Datenwissenschaft*), which aims to promote the transfer of knowledge within the Confederation and to encourage exchanges among the scientific community, research institutes, and bodies responsible for practical application, as well as to support the federal administration so that it can carry out projects involving data science.

On the health data side, following the analysis of the elements already in place and the challenges to be met in implementing personalized health and the Message FRI 2017-2020, the Confederation launched the Swiss Personalized Health Network (SPHN) initiative. The implementation of the initiative is led by the ASSM, the Swiss Academy of Sciences, which assumes organizational, legal, and financial responsibility for it. To promote personalized medicine and health in Switzerland, the aim of the SPHN is to create an infrastructure for networking the data of partner institutions, such as hospitals with research centers, universities, etc. Such an infrastructure will enable the interoperability of clinical patient data for research purposes.

At the European level, the European Open Science Cloud (EOSC) program supports the European commitment to transparent data-based science and to accelerate innovation. To this end, the EOSC will provide "*a virtual environment with open and transparent services for the storage, management, analysis and reuse of research data across borders and scientific disciplines by federating existing scientific data infrastructures"* for professionals in science, technology and humanities in the European Union.[5]

This is not a complete overview of all the different ongoing initiatives. Currently, the different activities and initiatives are not well coordinated, and a comprehensive national strategy to make data available

[5] https://www.eosc-portal.eu/about/eosc

for research purposes and how this links to European and international developments is not yet in place.

4 Practices and the legal environment in selected countries

The legal and institutional frameworks of other countries are elements for reflection for proposals to facilitate access and data linking for researchers in Switzerland. The chosen countries are Germany, France and Finland, as all three have developed effective solutions for accessing data for research in different ways and in different legal and institutional organizational contexts, offering a wide range of solutions. We also provide insights into the legal framework of the General Data Protection Regulations (GDPR), insofar as EU countries are subject to it.

4.1 Framework law for data access and linkage in European Union (EU) countries

The aim of the GDPR is "to guide data processing, increase trust, and encourage sharing and reusing data." For each of the countries of the European Union, the GDPR is articulated with the national laws, which may "maintain or introduce more specific provisions to adapt the application of the rules" (GDPR, chap I, art. 6; chap IX).

4.1.1 Data information and documentation (metadata)

The dissemination of information and documentation of data is governed by the "Data Transparency Directive" (EU Directive 2019/1024), which encourages (and can finance) the public sectors of member states to "make data easily accessible for reuse." In addition, the Commission plans to fund the development of interoperable Common Data Spaces in strategic sectors to overcome legal and technical barriers by promoting (i) the deployment of data-sharing tools and platforms; (ii) the creation of data-governance frameworks; and (iii) the improvement of data accessibility, quality, and interoperability. These spaces aim to encourage the sharing of information on data, including that produced by private companies (A European strategy for data, Brussels, 2.19.2020 COM (2020) 66 final).

4.1.2 Data access and linkage for research in the GDPR

Access to and processing[6] of sensitive data from the administration and the private sector for public research is only permitted by the GDPR (chap. I, art. 5) under certain conditions (chap. IV). Access to personal data is granted if the applicant's interest in the information outweighs the interest in protecting the individuals concerned or if they have given their consent. Irreversibly anonymized data, i.e., data that no longer allows the re-identification of a person, are not subject to the regulations on the protection of personal data, unlike pseudonymized data.

Moreover, to be authorized by the GDPR, accessing sensitive data requires the prior consent of the supervisory authorities (chap. IV, § 1, art. 31) and the organization of data security (chap. IV, § 2) in advance (chap. IV, § 1, art. 25). The processing of personal data must be documented by a data controller (chap. IV, § 1, art. 24) and recorded in a register of processing activities (art. 30) kept by a data

[6] In the GDPR, data linkage is included in data processing. Data processing is defined as any operation "applied to personal data or sets of data, such as collection, recording, organizing, structuring, storing, adapting or modifying, retrieving, consulting, using, communicating by transmission, dissemination, or any other form of making available, linking or interconnecting, limiting, erasing, or destroying" (chap. I, art. 4).

protection officer (DPO) (chap. IV, § 3, arts. 37, 38). The DPO is responsible for ensuring that the rules on the protection of sensitive data are applied by the data controllers (art. 39). The regulations on the protection of sensitive data and the security of their processing (chap. IV, § 2, art. 32) include "the pseudonymization and encryption of personal data," the means implemented to guarantee confidentiality, the "integrity, availability, and constant resilience of processing systems and services," the means to restore availability and access to data in the event of an incident, and a "procedure to regularly test, analyze, and evaluate the effectiveness of technical and organizational measures to ensure the security of processing."

The organization of data security also includes an impact assessment on data protection (chap. IV, § 3, art. 35) made by the data processor and a prior consultation of a supervisory authority on the aims and means of the proposed processing and the measures and guarantees provided for data protection (chap. IV, § 3, art. 36). This must contain at least "a systematic description of the envisaged processing operations and the purposes of the processing", "an assessment of the necessity and proportionality of the processing operations in relation to the purposes" ,"an assessment of the risks to the rights and freedoms of data subjects", and "the measures envisaged to address the risks, including safeguards, measures and security mechanisms to protect personal data."

Data storage: Personal data may be kept for a predefined period of time, fixed or according to a predetermined logic, and for a duration in accordance with the purposes of the research and/or the legal basis for processing (art. 13).

Reuse of data: Sensitive data may be reused, provided that the purpose of the further processing is compatible with the original research purpose (arts. 5, 14) and that appropriate technical and organizational data protection measures are implemented.

4.2 Legal and political framework in Germany

4.2.1 Institutional landscape for official statistics

Germany is a federal republic, and the official statistics system is organized accordingly. The statistical system is partly decentralized and is composed of the *Statistische Bundesamt* (Destatis), which is the federal statistical office, and fourteen Länder statistical offices, which are independent of the Destatis. Other institutions also produce statistical data. These include the Research Data and Service Center of Bundesbank (RDSC) as well as other research data centers (RDCs) accredited by the German Data Forum (RatSWD). The RatSWD plays an important role on many levels: it advises the national government on strategic issues and also accredits and evaluates data centers.

4.2.2 Data information and documentation

The German Freedom of Information Act (*Informationsfreiheitsgesetz*) provides for a right of access to information to the completed and documented data of public administrations (arts. 3, 4). This includes information related to personal data (art. 5) only if the applicant's interest outweighs the interest of the person to whom the data relates or if the latter has given his or her consent. It also does not apply to company data (art. 6).

The same type of laws has also been adopted at the Länder level.

4.2.3 *Legal regulations for administrative and private data access and linkage*

Data access: Administrative authorities under public law and federal authorities are obliged to provide administrative data as open data (art. 12 of the Data Transparency and eGovernment Act).

The Federal Data Protection Act (BDSG—*Bundesdatenschutzgesetz*) and the data protection laws of the Länder (LDSG—*Landesdatenschutzgesetz*) supplement the provisions of the GDPR that permit the processing of personal data for substantially predominant research interests (art. 27, para. 1), with the obligation to anonymize the data processed (§ 3, art. 27) and to pseudonymize or anonymize the data during processing (§ 1, art. 89).

Data linking: There is no legal basis in Germany for data linking for scientific purposes. This is governed by the GDPR, but with restrictions issued by the German Federal Statistical Law. These restrictions mainly concern the linkage of survey data with register data or the linkage of business data from the Federal Statistical Office with microdata from the Institute for Employment Research of the German Federal Employment Agency

4.2.4 *Institutional and technical provisions for data access and linking: Third-party centers to access and link sensitive data*

The federal statistical office Destatis and the Länder statistical offices mentioned above form a network of statistical centers that provide anonymous formal and factual microdata (3rd Federal Statistical Act, art. 16, para. 6, BstatG). They also provide social security codes, which can be used in the case of data linkage (SGB X, art. 75, allows the provision of data from social insurance schemes).

One salient point is that in Germany, access to sensitive data (and, to some extent, the linking of such data) is carried out through dedicated centers that are accredited by the German Data Forum (RatSWD). The RatSWD has acted as an advisory council to the federal government since 2004. It aims to improve the use of data to enable the construction of evidence-based policy and to facilitate access to high-quality data for research through Research Data and Service Centers (RDSC). The RDSC aims to improve the use of data by (*i*) ensuring access to sensitive data for academic research while guaranteeing the protection of these data, (ii) documenting data and methodologies for processing these data, and (*iii*) practicing record linking and linking of microdata, including company data, banking data, and statistical data.

There are currently 38 research data centers (RDCs) 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. Researchers must justify the need for their processing, travel to an RDC to process the data, and do all the analyses on a computer in the RDC. Half of the RDCs also provide data linking. The federal statistical law is applicable neither to all RDCs nor to all projects they carry out, as the legal framework of reference varies according to the type of data and their sensitivity.

4.3 Legal and political framework in France

Access to statistics produced by the public sector and administrations is governed by the articulation of the GDPR and four national laws: the CADA law (the Commission for Access to Administrative Documents), the Archives Law (which defines the regime of access to administrative documents), the Data Protection Act (relating to specific secrets and all personal databases), and the Digital Republic Law. These laws are articulated in the RGPD.

4.3.1 *Institutional landscape for official statistics*

France, as with its statistical and political system, is more centralized than in Germany or Switzerland. The official statistics system (*le service statistique public*—SSP) in France is coordinated by the National Institute for Statistics and Economic Studies (INSEE—*Institut National de la Statistique et des Etudes Economiques*). The INSEE, created in 1946, is a general directorate of the French Ministry of Economy and Finance. Coordinated by INSEE, the ministerial[7] statistical services also carry out statistical tasks relating to public administrations. There are no regional statistical institutes.

4.3.2 *Legal regulations for administrative data access and linkage*

The information produced by the administration is openly communicable (CADA law), except if it interferes with the exercise of the State's regulatory activities or with private life, medical secrecy, business secrecy, if it is not possible to conceal or separate the information, or if it contains secrets protected by law. However, various legislative and institutional measures taken in recent years have removed most of the legal obstacles to researchers' access to administrative data and have also facilitated the linking of administrative data.

Measures to simplify data access and linking were based on the creation of the *Comité du Secret Statistique* (Committee on Statistical Confidentiality), which ensures compliance with the rules of statistical confidentiality and gives its opinion on requests for the communication of individual data collected through statistical surveys or transmitted to the official statistical service. The Statistical Confidentiality Committee is chaired by a State Councilor and is made up of administrations, archives, researchers, and employers' and employees' unions.

With the help of the Committee on Statistical Confidentiality, these measures provide for the following:

- Access for research purposes to data on companies collected by the INSEE (French National Institute for Statistics and Economic Studies), with a derogation granted by the Committee on Statistical Confidentiality.
- A broadening of the possibility of a derogation of statistical secrecy for confidential public statistics data on persons and households upon a decision by the Archives following an opinion given by the Committee on Statistical Secrecy (Archive Act, 2008).
- The generalization of access for research purposes to all other administrative databases covered by professional secrecy (with the possibility of referral to the Committee on Statistical Confidentiality for these databases; Law on the Digital Republic, 2016).

Amendments and derogations have also been specifically promulgated for the various sectors of sensitive data (tax data, banking data, medical-administrative data, etc.). Specific provisions are provided for health data in order to cut down on the red tape related to the processing of such data for research while strengthening the protection of the data concerned.

[7] A *ministry* in France is a division of the central public administration responsible for implementing government policy in a specific area.

4.3.3 *Institutional and technical provisions for data access and linking: Third-party centers for the access and linking of sensitive data.*

To strengthen the protection of sensitive data and to inform individuals, institutions, and companies, France has created the CNIL (*Commission Nationale de l'Informatique et des Libertés*—National Commission for Information Technology and Civil Liberties).

The CNIL is responsible for ensuring the protection of both public and private personal data. It is an independent administrative authority (AAI—a public body that acts on behalf of the State without being placed under the authority of the government) created by the Data Protection Act to facilitate access to sensitive data while guaranteeing its security. The *Centre sécurisé d'accès aux données* (CASD—Secure Data Access Center) was set up in 2009 at the initiative of INSEE to provide researchers with access to official statistics data. CASD is a public interest organization that brings together the INSEE and also research institutions such as the CNRS, the *École Polytechnique*, and HEC Paris; it was created by an interministerial order on December 29, 2018.

The CNIL provides secure remote access for researchers—in France, in the European Union, and, under certain conditions, in North America—to all the very detailed data in official statistics as well as to a growing number of administrative databases, including tax and medico-administrative data. Remote access is conducted by sending a highly protected package containing the data and the software to be analyzed after the researcher has gone through various ethics committee checks and has received training in data security.

Data linking has also been facilitated by these derogations and is based on the involvement of trusted third parties, i.e., institutions that act as third parties between data owners and researchers.

For the linking of personal data that is not classified as sensitive, data can be linked using the NIR (*Numéro d'inscription au répertoire*), which is the common identifier for databases produced by the French public sector. However, researchers are not allowed to use the NIR and must go through two trusted third parties. The first (a section of the INSEE) oversees the recoding of the NIR and the generation of a new linking key. The second trusted third party, which must be a highly secure center, carries out the linking.

No specific institution is responsible for the entire process and, for example, the CASD, the secure data center, can act as a second trusted third party for researchers. The keys are specific to each project and must be destroyed at the end of the project period. Deterministic linking is not allowed for sensitive data. Probabilistic linking is possible, subject to the granting of access to data for research projects by the CNIL and the Statistical Secrecy Committee.

4.4 Legal and political framework in Finland

4.4.1 *Institutional landscape for official statistics*

In Finland, access and linking to official data is regulated by the Finnish Statistics Act (280/2004) and the GDPR. It is carried out through Statistics Finland (*Tilastokeskus*), the "only Finnish public authority specifically established for statistics."[8] As in France, Finnish regional statistical production is centralized and carried out by Statistics Finland. Statistics Finland provides information on the statistics it produces as well as on the statistics produced by the 11 other public organizations producing statistics.

[8] https://www.stat.fi/org/index_en.html

4.4.2 Legal regulations for administrative data access and linkage

The Finnish Statistics Act allows access to microdata for research. Some microdata is available through Statistics Finland. Access to other data is granted through *Findata,* which is the "Health and Social Data Permit Authority" that operates under the guidance of the Ministry of Social Affairs and Health.[9] Since April 1, 2020, it has also been possible to access pseudonymized sensitive personal data for research purposes. The request to access sensitive personal data is made in the Findata system. If the request is accepted, Findata then takes over the collection of the register data from the different data owners, the pseudonymization and anonymization of the data, as well as the linking.

Findata provides a remote access environment for the processing of pseudonymized personal data: the researcher can thus analyze these data in a secured remote-access environment on a virtual machine using various types of standard statistical software, such as SPSS, Stata, SAS, and R. In order to use this secure environment, *(i)* researchers must have a license approved by Findata, *(ii)* they must have completed the remote access environment use form, *(iii)* the data must have been collected from the controllers, and *(iv)* an agreement between Findata and the researcher on the use of the remote access environment must have been made.

4.5 Differences between Switzerland and other countries

Germany and Switzerland, which are federalist countries, have statistical offices at the federal and cantonal levels, while France and Finland have only national offices.

Germany, France and Finland have set up third-party centers for secure data access in addition to their statistical offices. This is currently not the case in Switzerland.

In Germany, these third-party centers are the Research Data and Service Center of the Bundesbank (RDSC) and the research data centers (RDCs), which are part of both research institutions and government organizations. Sensitive data can only be accessed securely on site.

In France, the CASD offers secure remote access. Like the German RDCs, the CASD brings together the INSEE and research institutions. It was created on the initiative of the INSEE. In Switzerland and Finland, access to the data is through the bodies in charge of statistics.

In Finland, Findata has been established recently to grant access to public data, which also includes linked data.

In France, ethics and data security commissions have strengthened the data access procedures such that access to data is subject to the opinion of the CNIL and the authorization of the Statistical Secrecy Commission. In Switzerland, Germany, and Finland, the legitimacy of access to data is assessed by the bodies providing the statistics. Finland has the particularity of issuing data access permits.

[9] https://www.findata.fi/

	Switzerland	Germany	France	Finland
Organization of Statistics	Federal/National level: - Federal Statistical Office - Other federal offices and executive body of the Confederation Regional level: - Cantonal statistical offices	Federal/National level: - *Statistisches Bundesamt* - German Data Forum partially as regulator Regional level: - Länder statistical offices	Federal/National level: - National Institute of Statistics and Economic Studies (INSEE) - Ministerial statistical services No regional level	Federal/National level: - Finnish Statistical Institute Findata No regional level
Data linking centers	Federal Statistical Office	Research data centers (RDCs)	A section of the INSEE as the first trusted third party and other secure centers such as a CASD as the second third party	Findata
Third-party center for secure access		Research data centers (RDCs)	CASD (Secure Data Access Center)	Findata
Commission for Ethics and Secure Data Access		Committee on Statistical Confidentiality (CNIL), National Commission for Information Technology and Civil Liberties		Findata—Health and Social Data Permit Authority issued by the Findata data center

5 How to improve access to data for research

Switzerland does not yet have a general strategy to promote and facilitate open data access for research. Experience from other countries shows that it is possible to implement such a general open-access strategy with a simultaneous increase in the overall level of data protection. Privacy protection does not need to prevent or hinder the use of personal administrative or private data or the linking of data for research. Rather, it should guide the creation of a comprehensive legal and institutional framework that allows making data available for research while ensuring the protection of individuals' privacy.

Improving access to administrative and private data requires an *in-depth dialogue between researchers, research institutions, public and private data providers, as well as policy makers*. The value of access to good data for good research needs to be stressed more strongly in public contexts and in exchanges with the many different stakeholders. Much of the discussion around open research data is still centered on making data produced by researchers FAIR. The dialogue does not expand to making other types of data FAIR. Key concepts established around research data, such as the FAIR principles

and their implementation, research transparency, and reproducibility, are not yet established outside the academic world. Further, researchers often fail to understand the limitations and procedures necessary to access administrative and private data for research. The establishment of a network of partners for dialogue would also facilitate the sharing of a common understanding, best practices, as well as technical expertise on how to make data available and how to link data.

This dialogue must simultaneously address the key principles of open data and research transparency as well as data protection and privacy rights. While there is a demand for more open data, there is, at the same time, a demand for more data protection. Data access must be regulated: there is an increasing and legitimate awareness of the need for the privacy protection of individuals in a digitalized world.

Allowing access to data for research does not imply that non-anonymized personal data should be shared. On the contrary, personal data should be protected at all cost, and the risk of disclosure needs to be minimized. With appropriate procedures and institutions in place, it is possible to improve both principles simultaneously. This last section of the report outlines some key principles to help envision what this could look like. We propose several concrete institutional and technical avenues (5.1), proposals for changes to the legal framework (5.2), as well as a proposal for next steps (5.3).

5.1 A favorable institutional framework

There are several building blocks of an institutional framework. In principle, many of the following building blocks of an institutional framework are relevant for data producers from the private or public sector and also data produced by researchers. And some regulation, for example with respect to data protection, is relevant for all personal data. However, variation exists in different ways:

- The administration can only collect and process data if a legal basis exists. This is not the case for private and research data, which usually collect and process data without a specific legal basis.
- Whether data access should or can be a requirement depends on the identity of the data producer and for what purpose the data has been collected. There is a mixture of public versus private interest why certain data should or should not be accessible in different ways. Data access to research data is required by research funders. Under open government data principles, administrative data may also be accessible in principle, however, data protection or state interest may limit access. Data from private companies is currently seen widely as private property. However, the current pandemic or the debate on the influence of social media on democratic processes may in the future also lead to reflections on whether under some circumstances there is a public interest that such data in the possession of private companies should be publicly available.
- Some processes are clearly established in some fields, for example the requirements of ethical approval in human research, while other fields do not require such an approval.
- How sensitive data is depends on the information it contains and on the specific context, and needs to be assessed on a case by case basis. Especially sensitive are information on religious, ideological, political or trade union views or activities, information related to health, privacy or racial affiliation, social aid, as well as administrative or criminal prosecutions and sanctions. However, depending on the specific context or whether information is combined, other data may also be worthy of protection. Data protection requirement then also depends on how widely such information is available. Information on an entire known population that is non-anonymized may need to be more protected than information from anonymized small samples where it is very difficult to identify individuals in practice.

5.1.1 Metadata, procedures, and guidelines

The first requirement to access and link data is *metadata*:

- *Rich metadata*. Often it is not clear what data exist and/or what the structure and content of existing data look like. Without better and rich metadata and documentation, the potential of existing datasets is hard to unlock. To make metadata useful and exchangeable, it should follow some common standards. This is true for all kinds of metadata. With respect to administrative and private data metadata, these should also include information on existing personal identifiers that potentially enable the linking of different datasets as well as provide information on how the data access is regulated.
- *Making metadata findable* in publicly accessible repositories: While it is likely that not all metadata can be stored in a single system, it makes sense to try not to multiply metadata catalogues. A central catalogue for the entire federal administration and, possibly, also for cantonal data, where metadata is stored from different national and cantonal data producers, is desirable. The centralization of systems makes it easier to develop and implement standards, and linking different metadata catalogues also becomes less costly if the number is low. Several initiatives are already moving in this direction within the academic domain, including the Connectome initiative, whose objective is to set up a platform to access information on research data wherever it is stored. This information is related to the datasets, such as where they are stored, but also to the other datasets of projects in which the data have previously been used and the publications in which they are cited.

Transparent and standard *procedures and guidelines should be in place pertaining to how data can be accessed, linked, used, and stored*. There needs to be regulation in the following areas:

- In principle, *who can get access,* and for what purpose? Open sharing of government data is already a key principle in Switzerland; however, the open data principle does not extend to non-anonymized personal data. This needs to remain in place. Non-anonymized personal data should not be released as all residents have the right to expect that their data will remain protected. Nor should personal data be used for regulatory purposes, e.g., to control or target individuals or companies. However, anonymized personal data should be usable for research purposes in a comprehensive way.
- *Who decides,* in specific cases, on access and addresses ethical, legal, or data protection concerns? Because data may be sensitive, there always needs to be an assessment of whether the value of a research project outweighs any potential harm. For research on humans, this is regulated by the Human Research Act and delegated to cantonal ethics commissions. Many universities also have ethical boards in place that grant ethical approval that, in some cases, is not covered by the cantonal ethics commission. However, these structures cannot easily be used for all administrative or private data. Therefore, a special body needs to be created to assess potentialities and risks when granting access or allowing data linking.
- *What are the contractual procedures to access data*? While contractual agreements may be necessary in many cases, it should be possible to organize this in a short amount of time within a general contractual framework that does not require complicated formal approval processes for each individual project.

- *How is access organized?* In most cases, only access to anonymized data may be granted, but entire datasets can be transferred to researchers. However, anonymized data also bears the risk of identification of individuals, especially if a dataset contains a great deal of information. Therefore, data access procedures may also vary depending on the sensitivity of data (see below).

5.1.2 Key principles for linking data and accessing linked data

While comprehensive access, use, and storage are relevant for all kinds of data, linking data requires specific procedures for two reasons. First, linking datasets is not possible without identifying information. Second, the sensitivity of data and, hence, the potential harm may increase once data is linked.

To address these concerns, specific procedures can be put in place to link data. This should be based on two different principles:

The separation of identifiers and substantive variables and the separation of the process of creating a linking key from the process of merging data with a new key. For research, this is not problematic since no substantive information is lost:

- For the linkage process, only the identifying information (identifiers, name, date of birth, etc.) is required, not the other variables.
- For data access, the linkage key is the only way to merge the data without any identifying information being necessary. Researchers do not need identifying individual data as long as personal data can be merged.

Separating the different processes greatly reduces the risk of disclosure and data breaches of linked data because identifying information and substantive information from different data sources are never in the same place. This solution is applicable to research data as well as to administrative or private data. This security guarantee could help in persuading companies holding sensitive data to share it.

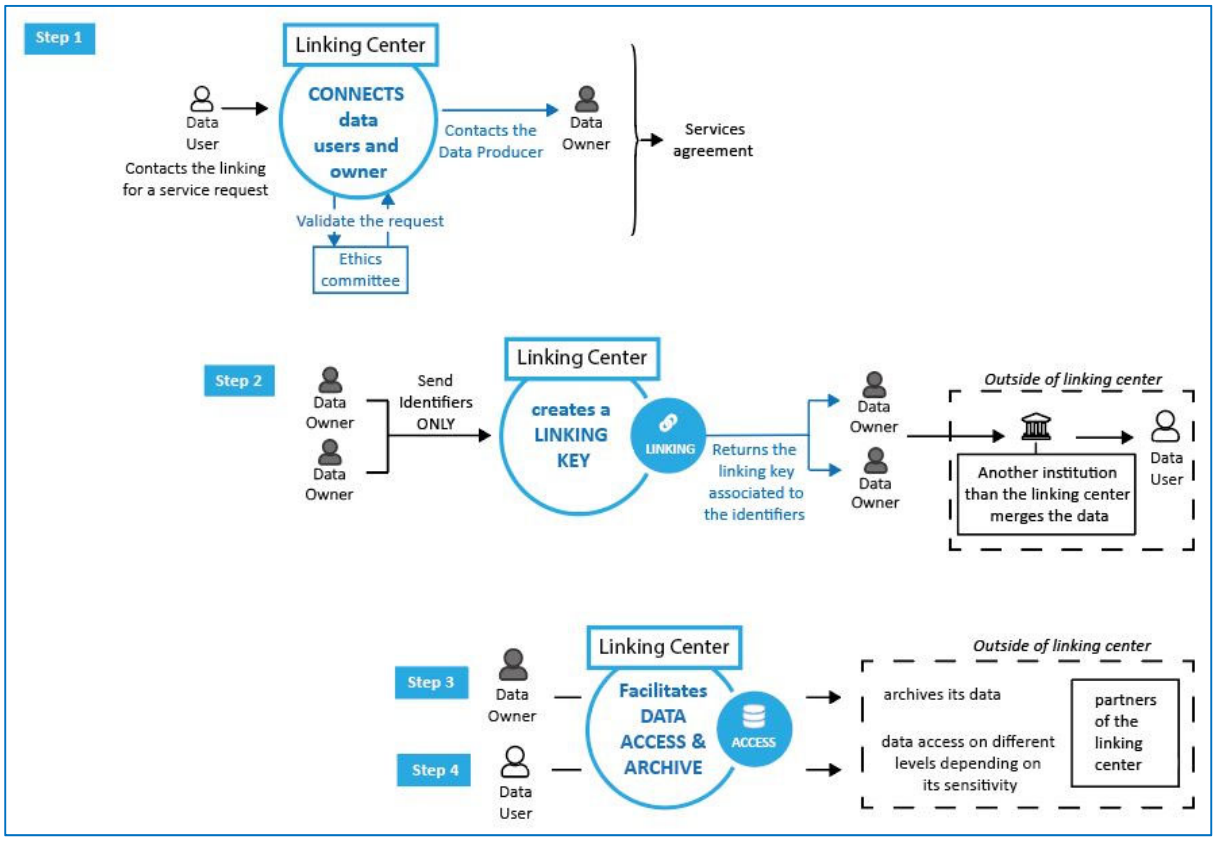
To implement the separation of identifiers and variables, the use of a trusted third party and a specific *data linking center* appears relevant (as exists in France, Germany, or Finland).

Such a linking center would handle different steps of the process and could look like this (see Figure 1):

- In *step 1* a user makes a request and the linking center acts as an intermediary between data users and data owners. The center handles the project approval process including ethics committee approval if required and handles the contracting of the linking/service part. The center could also handle the secure transfer of data, regulation and certification of access points, as well as providing knowledge on data linking and how to use linked data more generally.
- In *step 2* the production of a linking key takes place. A linking center would generate the linking key from the identification variables without ever having access to the information variables that the researcher wishes to link. The different steps are as follows:
  - The data owners send only the identifying information to the linking center, without the other substantive information/variables.
  - The linking center links the different datasets using unique identifiers or through probabilistic matching and generates a new linking key.

- The center returns the linking key associated with the identifying variables of the original data to the data owners, and the data owners add this linking key to the original dataset. The data owners remove all identifying information and send the dataset to the data user or to the institutions granting access to the linked data, potentially in a secure environment.
- The data merging, i.e., the actual assembly of the datasets using the linking key, is carried out elsewhere by the data owner's institution or another partner institution specializing in data management with a secure infrastructure.

Figure 1. A linking center: key processes



- *Step 3 and 4* focus on data archiving and data access to linked data. This is done outside the linking institution but may be under the guidelines and regulations of the linking center. Access at different levels depends on the sensitivity of data and contracts: A complementary solution to enhance data security is to implement differentiated access according to the sensitivity of the data: on-site access for highly sensitive data, secure remote access for sensitive data, and download access for less-sensitive data.
  - *On-site access*: For highly sensitive data, the data can only be analyzed in a protected environment within trusted secure centers, where no copies of the data can be made and only the results tables of the analyses sufficiently aggregated to comply with statistical confidentiality can be taken away. Such secure centers have already been developed in various countries: in Germany with the RDC, in France with the CASD, in the USA with the pioneering Cornell Restricted Access Data Center built in the 1980s, and in Canada. For the future, it is also possible

to think about the establishment of a network of secured centers hosted by data providers or within academic institutions such that a secure center allows access not only to data from one center but from other centers simultaneously.

- *Remote access, remote desktop or remote analysis*: Data can be analyzed but not downloaded. Remote access has been experimented with in several countries: in France or Finland, as described in part 4, and also in the United Kingdom, Denmark, the Netherlands and Sweden. The modalities of remote access, although differing according to the legislation of the respective countries, all follow the same principles: the researcher has to install remote-access software that allows a connection to the server hosting the data via a secure channel, and the data cannot be downloaded.
- *Data downloads*. Less sensitive or irreversibly anonymized data can be distributed by the data owners or data archives, as is the case already. Typically, users have to sign data contracts wherein they commit to protect the confidentiality of respondents and that they will not use the data for anything but research. This is the current practice not only for existing data archives but also for public data distributed by the Federal Statistical Office.

Institutionally, such a linking center could either be a separate legal entity or embedded in an existing institution. It could, for example, be an annex of a statistical office (such as Findata in Finland) or a stand-alone structure (such as the RDCs in Germany or the CASD in France). To attach such a center to an existing institution has the advantage that it can rely on the institution's experience of how to handle data. An independent trusted structure might, however, provide greater incentives for the different data producers from the government, the private sector and research sectors to make their data available to third parties. Some data owners might be reluctant to send their data to a center that is fully part of the central administration. It is also imaginable that different centers of this type might exist, one to handle linking and access to administrative data at the national level, and one that handles linking and access to data that includes data from other sources, for example, from researchers or private companies.

In any case, *building an infrastructure network* to include these different potential centers and the archiving platforms would be necessary to ensure the proper execution of all the processes and data transfers and that sensitive data is not stored on private infrastructures (e.g., of companies), publishing houses, or journal websites, and that privacy and data rights and intellectual property rights are protected within Switzerland. Private infrastructures located outside the country follow different protection rules, which is especially problematic for sensitive and linked data. The construction and sustainability of such an infrastructure must be based on a sustainable economic model, to be defined.

5.2 A favorable legal framework[10]

The future legal framework should govern access and data linking of sensitive data for research purposes. This should include the following parts:

- Access to all government data for research should be granted in principle.

[10] As a legal analysis, this part is also largely based on an outline that FORS commissioned from the legal services of SWITCH (Anna Kuhn and Nora Zinsli).

- Policy makers should consider that access to data for research should also be extended to private data if there is a predominant public interest. This could, for example, be for research around a health crisis or for research into the functioning of the democracy.
- A legal framework following the principles outlined in chapter 4 should specify under what conditions and how access to data is granted.

While there is a need for a more in-depth analysis of the legal situation at the national and the cantonal levels, first reflections point in the direction that it could make the most sense to insert a legal basis for data access for research purposes into the Federal Act on Freedom of Information in the Administration (FoIA, *Öffentlichkeitsgesetz*; see chapter 3.1). The FoIA covers all activities of the federal administration and public bodies and is therefore broad enough to enable a general approach to granting access to data. The FoIA restricts access to certain information where there is a public interest in its secrecy (FoIA, art. 7). Furthermore, the FoIA already includes a procedure for data subjects to exercise their rights. The procedure in the FoIA should, however, be adapted to the specificities of the exchange of large amounts of data for research purposes. The new section in the FoIA should, therefore, regulate access to federal databases for research purposes and regulate the safeguards of public and private interests along the time axis of the processing of this data from the moment of access, through the processing and research activities, to its publication. Furthermore, since the Federal Statistical Act already has a provision on linking data, this provision could also extend the role of the Federal Statistical Office to making data available to third parties for research purposes, which currently fall outside the Statistical System.

While these two laws could represent a solution for data through the central administration, the question of how access to other data — whether from the cantons, from research, or from the private sector — remains open and calls for further analysis. Special attention should be paid to the question of the applicable laws during the actual research work and the publication of the research results. Once access to databases has been granted, certain research institutes are subject to cantonal law—universities and higher education institutions are particularly worthy of mention. As a result of this fact, especially in the area of data protection, data subjects cannot exercise their rights centrally vis-à-vis an authority under federal law but would have to assert the relevant rights in the relevant canton to the relevant competent authority or institution in the relevant official cantonal language. This would make it very difficult or practically impossible for data subjects to assert their rights. We therefore see the necessity of examining the idea of a centrally administered office that can receive and process inquiries in all official languages of the Swiss Confederation and monitor compliance with Swiss law by research institutes. However, this may lead to a conflict with the interests of the cantons, which would have to accept the applicability of federal law and the supervision of data processing by a federal authority.

Adding further to the complexity are research bodies governed or sponsored by international or multinational laws. Research institutions are increasingly funded by inter- or multinational organizations and/or conduct work in a decentralized fashion, scattered over institutions based in several countries. In this context, the existence of a framework for the envisioned international exchange of data for research purposes should be examined. In this context, bi- and multinational agreements seem necessary in order to define an information security and data protection standard that research institutes should meet when exchanging data with international partners for the purpose of their research.

5.3 Next steps

This report outlines various issues and challenges in relation to data access and data linking. In order to improve access to data for research in the near future, various further steps are necessary. In order to define and implement a strategy and to steer this process, a working group with different stakeholders from scientific institutions and organizations and the federal and cantonal administrations could be put in place. Concretely, the following next steps would be useful:

- *A joint strategy.* Key academic stakeholders, such as Swiss Academies, the SNSF, and swissuniversities, as well as the SERI and other Federal Offices should recognize the importance and necessity of a joint strategy to facilitate access to data for research in a comprehensive way that extends beyond data produced by researchers. This is essential in Switzerland to ensure the competitiveness of the Swiss research system. They should also take the lead in moving such a strategy and its implementation forward.
- *Metadata*. Rich metadata is essential as the basis for data access and data linking because if it is not clear what data is available and where, it is impossible to use this data for research. At the national level, the Federal Statistical Office is about to provide metadata for all data collections at the federal level. For research data, metadata should be provided by the respective data archives. How metadata for cantonal data collections and also for private data collections would be organized remains an open question. In addition, it would be important for metadata to be findable across different metadata providers; therefore, metadata providers would need to ensure interoperability.
- *Institutional framework.* Proposals for an institutional framework to access and link data need to be further developed. While some key principles for an institutional framework are outlined in this report, many questions need to be addressed in greater detail: for example, who would govern and fund such a center and how many aspects of the process for data access and linking would need to be organized.
- *Legal framework.* Preparation of concrete proposals to amend the existing legal framework to grant access to and link not only administrative data but also, under some conditions, private data as well as research data should continue. This requires further analysis of the national and cantonal legal frameworks and should address questions such as how research institutes currently and in the future comply with existing laws while carrying out their research (i.e., after they have been granted access to data), what law(s) are applicable to the research institutes (international, federal, cantonal) while processing data, what bodies make decisions regarding the access to data, what body or bodies would supervise processing activities and data access, how and against what institutions should data subjects be able to make use of their rights, what standards should be applicable in international settings, and what would the Confederation do in order to ensure an adequate level of safeguarding of public and private interests.

6 Resources

Relevant legal texts

- Federal Data Protection Act: https://www.admin.ch/opc/de/classified-compilation/19920153/index.html
- Federal Act on Freedom of Information in the Administration (FoIA)/Öffentlichkeitsgesetz: https://www.admin.ch/opc/de/classified-compilation/20022540/index.html
- Federal Statistics Act (FStatA): https://www.admin.ch/opc/de/classified-compilation/19920252/index.html
- Federal Act for the Promotion of Research and Innovation and the Act on Human research: https://www.admin.ch/opc/fr/classified-compilation/20091419/index.html
- General Data Protection Regulation (GDPR)—Official Legal Text: https://gdpr-info.eu/

Germany

German Freedom of Information Act (Informationsfreiheitsgesetz):

http://www.gesetze-im-internet.de/englisch_ifg/index.html

Federal Data Protection Act (BDSG—Bundesdatenschutzgesetz):

https://rm.coe.int/09000016806af19d

Federal Statistical Act:

https://www.gesetze-im-internet.de/bstatg_1987/

France

Commission for Access to Administrative Documents—Loi CADA

https://www.cada.fr/

https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000020566920/2009-05-01/

Archives law—Loi sur les archives

https://francearchives.fr/fr/actualite/44101

https://www.senat.fr/rap/a07-147/a07-1479.html

Data Protection Act – Loi sur la protection des données

https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000037085952?r=F5TlyaWmap

Digital Republic Law—Loi pour une république numérique

https://www.legifrance.gouv.fr/loda/id/JORFTEXT000033202746/2020-10-06/

Site de la CNIL: https://www.cnil.fr/

https://www.vie-publique.fr/eclairage/19591-protection-des-donnees-personnelles-essentiel-loi-cnil-du-20-juin-2018

Finland

Finnish Statistics Act (280/2004):

http://www.stat.fi/meta/lait/statistics-act-2802004_en.html