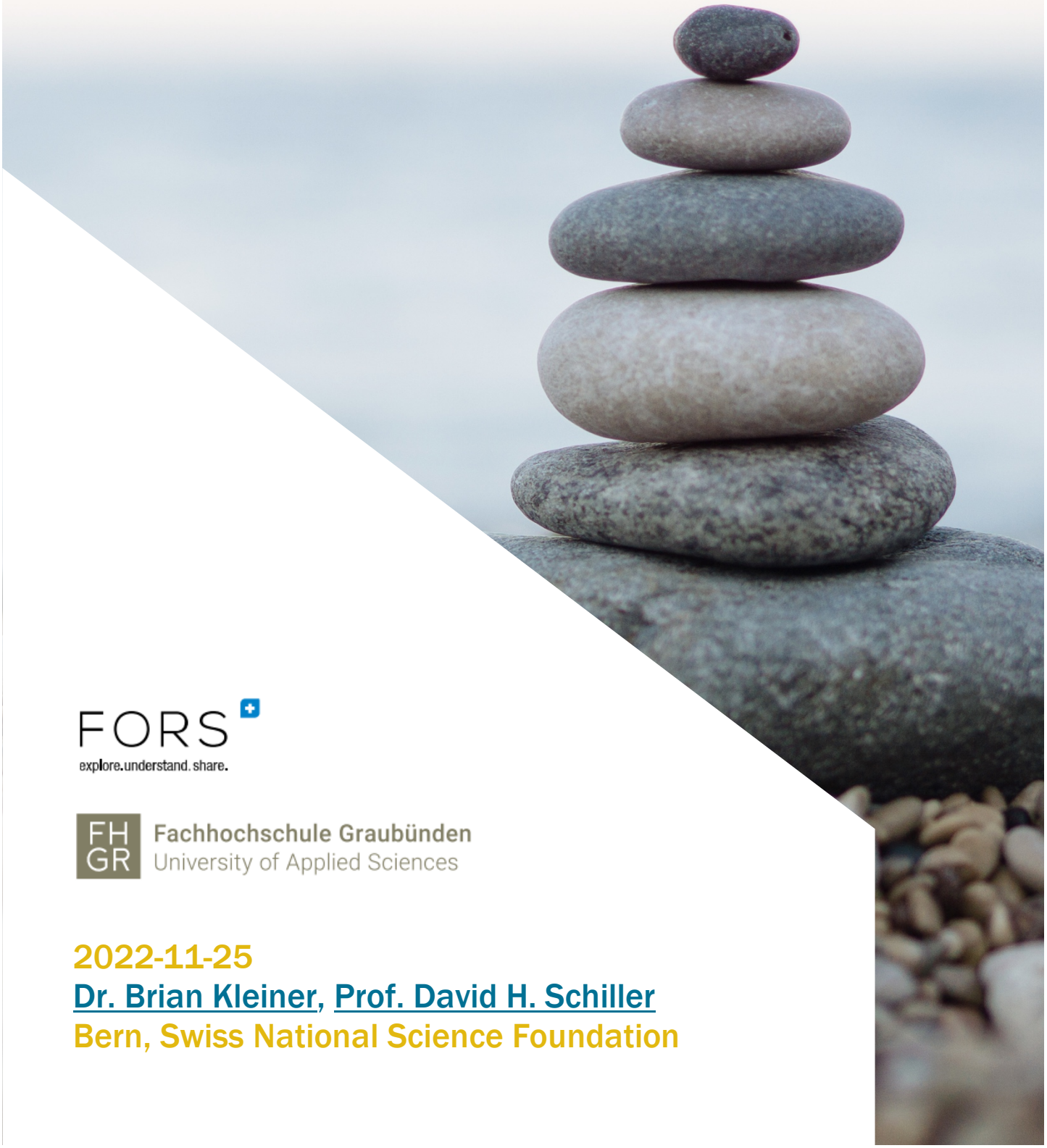
Workshop «Safe access to sensitive research data»



FORS
explore.understand.share.

FH GR Fachhochschule Graubünden
University of Applied Sciences

2022-11-25

Dr. Brian Kleiner, Prof. David H. Schiller
Bern, Swiss National Science Foundation

Summary

The first part of the workshop discussed and highlighted various international approaches to ensure safe access to sensitive data. Whether physical safe rooms, federated data analysis systems, microdata services through an API and scripts, specific hardware boxes with fingerprint identification or a network of guest researcher workrooms, the solutions are manifold and developed out of specific needs.

The second part of the workshop gave insight what personal data really is, how data linkage in Switzerland has evolved, which data access challenges the health sector faces and provided an overview of the educational trajectories LABB study project.

In the concluding discussion round its was agreed that despite all differences between the different sectors and other unsolved problems (money, preservation) there is a common need for a coordinated, specialized, and shared infrastructure and an alignment on strategies, governance, and standards.

Keynote 1

Dr. Deborah Wiltshire / GESIS, Germany

Deborah manages the Secure Data Center at GESIS which provides access to data requiring special protection.

- *There are 3 types of data:*
    1. *Identifiable data*
    2. *Pseudonymized data (re-identification possible)*
    3. *Anonymized data*
- *Safe rooms are still dominant as today:*
    - *Access is controlled, thin clients*
- *Remote access is coming more and more:*
    - *Still based on the safe room concept, but the safe room is at the partner organization*
    - *Connection via encrypted internet connection from the own office -> you lose many of the direct physical controls if the safe room is not in your organization anymore*
    - *Alternative is "Remote execution" -> researchers don't work with the raw data, the get only the results*
- *Human model of data security:*
    - *Most mistakes happen on the level of humans which are involved in the service*
- *GESIS has incorporated the 5 Safe Framework:*
    1. *Safe Projects:*
        - *Happens at the application stage -> Use of data agreements, very specific*
            - *Who wants access, what is the purpose, how long should the access be granted, what are the terms/conditions*
            - *Sometimes institutional agreements*
        - *Applications are assessed (mostly by data owner itself)*
            - *Is it feasible*
            - *Is the statistical purpose valid*
            - *Is it for public good*
        - *Applications are checked case by case or by a data committee (group of experts, assessing is done once a month)*

- *Good example for a safe project is METADAC[1] (on the site you will also find a lot of great resources for managing data access)*

2. *Safe People:*
   - *Training for researchers is crucial, in UK it is mandatory*
   - *The Safe Data Access Professionals (SDAP) Network and Social Science & Humanities Open Cloud (SSHOC) websites have a lot of good training resources*
3. *Safe Data:*
   - *Data must be confidential and treated securely*
   - *Removing direct identifiers is not sufficient, indirect identification is still possible*
4. *Safe Settings:*
   - *Thin Clients*
   - *Unique logins for researchers*
   - *Virtual Desktops -> no access to other clients/servers, the environment is sealed*
   - *Use of monitoring software*
   - *Safe rooms*
   - *Remote access:*
     - *2-factor authentication*
     - *Fixed IP*
     - *Training for researchers*
5. *Safe Outputs:*
   - *The goal is to minimize the risk ("We have taken all possible measures")*
   - *Statistical Disclosure Control (See for handbook SDC)*
     - *4 eye approach*
     - *Matches closely good research practices*

- *All 5 Safes together enables "Safe Use"*
- *Weakness of 5 Safes Framework:*
  - *Lacks the aspect "data stewards" (person who is controlling the access)*
    - *Often have little to no experience*
    - *Often no formal training for this job*
    - *They have to rely on a good functioning team*
    - *GESIS has resources for training*

[1] As a side note, if you want to know more about METADAC, you can also contact Rebecca Wilson (DataSHIELD), as she was involved in the development.

- *Challenges:*
  - *There is a growing amount of data*
  - *There is need for more linkage possibilities*
  - *New data forms -> DVD, social media (no legal framework)*
  - *Dissolving boundaries (Interdisciplinary & International)*
- *Nevertheless, of all measures -> TRUST is key!*

Session 1: International models and perspectives

Dr. Rebecca Wilson / DataSHIELD, UK

Becca is Principal Investigator and Lead of the DataSHIELD Project.

- *Goal is to find a balance between these 3 aspects:*
  - *Real risk of disclosure*
  - *Real costs of implementation*
  - *Real impact for researchers*
- *DataSHIELD is an open-source project*
  - *Federated analysis of data sets*
  - *Principle: "We take the analysis to the data"*
  - *It's not possible for researchers to view data on the individual level, but they can connect to all data sources for analysis*
  - *Study level & individual patient data meta-analysis possible*
  - *Established for longitudinal studies with the use of secondary data (topics are "hospitals" and "COVID")*
- *Inbuilt disclosure mitigations including Statistical Disclosure Control (SDC)*
  - *Compatible with formal data agreements, e.g., 5 Safes*
  - *Robust hardware, frequently updated*
  - *For study purposes no access to the primary live database*
  - *Only valid functions possible to run in a R parser*
  - *11 disclosure settings to check when generating outputs -> see https://bit.ly/DS-function-checks*
  - *Disclosure setting thresholds can be customized (e.g., only access for 3 users simultaneously)*
  - *All user commands are logged*
  - *Data controllers can block functions anytime*
  - *Open-source community -> maintains/updates the packages*

- *Challenges:*
  - *High user demand with attitude "taking and not giving"*
- *Strategies:*
  - *Train the trainers -> https://datashield.org/forum*
  - *Paid support for Community members*

Ørnulf Risnes / Sikt, Norway

Ørnulf is area manager for the microdata.no service at Sikt.

- *Microdata.no is a service for linkage and analysis of register data / other forms of microdata*
  - *Since 2018*
  - *Developed and operated by Sikt and Statistics Norway (SBB)*
  - *Full population data since 1964*
  - *Browser-based & location-independent*
- *Data & results are anonymous on the user side (built-in privacy controls)*
- *There is a regulation form, but it's very liberal for researchers to get access*
- *In the appendix of a research paper is no data anymore, instead only a script -> this can then be run by microdata.no & the output is then the same data as the original researchers has used*
- *Built in is also the 5 Safe framework -> but they changed "Safe projects" with "Safe purposes", as they don't deal with projects*
  - *"Safe settings" is central*
    - *Controlled execution -> own language and clients*
    - *API in front of secured data settings*
  - *Safe outputs*
    - *Noise added to all frequencies*
    - *Outliers removed*
- *Plenum questions ->*
  - *«Isn't it too limited? »*
    - *«We will support more and more data & variables.»*
    - *«Also, more and more statistical methods are possible.»*
    - *«Lend-out services are possible.»*

Kamel Gadouche / CASD, France

Kamel is the director of the French research data centre (CASD) public interest group which organize and implement secure access services for research.

- *CASD act as a thrusted third party for secure access*
  - *All the data & servers are in a secured bubble behind a firewall*
  - *The researchers/universities have remote access to the data through a specific device called SD-Box*
    - *Enrolment session with fingerprint authentication*
- *The data providers & users sign contracts with CASD*
  - *CASD does the data preparation & metadata catalogue integration*
  - *More than 700 projects and 1500 users France and Europe*
  - *247 data sources*
  - *More than 500 access points*
- *The framework for secure access & use of files is according 5 safes*
  - *Safe projects and safe people are checked by a committee of experts*
  - *Use of trusted certifications & security audits*
- *Publications based on the data go through a process where it generates a code for the researcher and the reviewer -> must match*
  - *Only the results are public, data can only be accessed through the SD-Box*
- *Plenum questions:*
  - *«Where is the SD-Box produced? »*
    - *«Inside the company. »*
    - *«There is also possibility to install virtual desktop environment.»*

Neil Murray / RDCnet, Germany

Neil is research associate at the Socio-Economic Panel (SOEP) since April 2021 and is part of the KonsortSWD team which establishes a research data infrastructure network (RDCnet).

- *Idea behind RDCnet -> Integration of different guest researcher workrooms (GRW) into a network of secure access points*
- *There are 4 components:*
  - *A contractual framework which serves as a base of trust:*
    - *Includes legal obligations with multilateral cooperation agreements*
    - *Physical safety standards*
    - *Data quality (only accredited RDC can take part)*
  - *Technical foundations which are:*
    - *Decentralization*
    - *Remote access (standardized software)*

        - *Role allocation (both for data receivers & data providers)*
        - *Flexibility (data providers can decide which data is shared and when)*
    - *Administration:*
        - *E.g., is the development of a joint platform where you can see all access points*
    - *Support:*
        - *Guides of implementations for a GRW and for implementing remote access*
- *Plenum questions:*
    - *«Is there also a possibility to access the data from home?»*
        - *«No, only from data safe rooms inside the GRWs?»*
        - *«More access points are possible in the future, e.g. inside universities, but we do not follow this presently.»*
    - *«How are you funded?»*
        - *«The German National Research Data Infrastructure (NFDI) is our funder.»*

Keynote 2

Prof. Dr. Georg Lutz / FORS, Switzerland

Georg is the director of FORS and professor of political science at the University of Lausanne.

- *There will be a new data protection law in place in 2023 -> definition of highly sensitive data expanded*
- *Data linkage mostly done by SFSO -> as researcher you can have data, but it is rather a complicated process*
- *A new (research) data policy should include:*
    - *An institutional framework providing metadata, data protection, linkage and regulations (permissions, data contracts, re-use)*
    - *A legal framework which regulates data use and linkage on the legal level*
- *There are federal efforts to improve access -> "Nationale Datenbewirtschaftung"-> new interoperability platform for Swiss administrative data*

- *Also, a new motion is discussed for the development of framework law for the secondary use of data*
- *But more work has to be done*
    - *Further cooperation (especially with SFSO) on how to make data usable und how to link*
    - *Conceptual work (see also report "Accessing and linking data for research in Switzerland") -> processes, funding, governance, institutional models*
    - *Public debate*

Session 2: National perspectives

Dr. Pablo Diaz / FORS, Switzerland

Pablo is strategy officer at FORS and holds a PhD in Social Science from the University of Lausanne.

- *Personal data is any information that relates to an identified or identifiable living individual*
    - *This also refers to information which indirectly identifies a person, e.g. a combination of identifiers or if you may need additional information which are freely available to be able to identify the person*
- *Laws are exhaustive, that means if it's not written in the law, it's not personal data, e.g. salaries are not sensitive data by the law*
- *Any data collection must have a purpose and this purpose must be evident to the data subject*
- *A layered approach is recommended:*
    - *Get consent*
    - *Anonymise the data*
    - *Implement access control mechanisms*

Dr. Ilka Steiner / BSV, Switzerland

Ilka is research project manager at the Federal Social Insurance Office (BSV).

- *Data Linkage Project -> Immigration and Settlement of Germans in Switzerland (Statistic of foreign nationals, household register, structural survey)*
    - *Before 2014 you could do the data linkage at the university and store the data on your own computer, but data access was very time consuming*
    - *After 2014, the linkage was only possible at SFSO*
    - *Now there are several data protection measure in place:*
        - *Contract*
        - *File transfer system (password)*
        - *Data storage checklist (local secure environment)*
        - *Methodological constraints for publication*
        - *Signed confirmation of deletion*
        - *Safe room access for data preparation, but not for data analysis*
    - *Plenum discussion:*
        - *Still not solved the needs between deletion of data versus reuse of data*

Dr. Sabine Österle / SPHN, Switzerland

Sabine oversees and manages the Personalized Health Informatics (PHI) portfolio and related activities towards achieving the interoperability goals in the Swiss Personalized Health Network (SPHN).

- *4 Data types in health sector:*
    - *Clinical data*
    - *Research data*
    - *Molecular data*
    - *Health citizen data*
- *To make this data "fair data" is hard, because a lot of the data consist of handwritten protocols by doctors*
- *How is the security handled?*
    - *Confidentiality -> only permitted persons have access (passwords, physical constraints)*
    - *Integrity -> unauthorized alteration is prevented (hashes, digital signatures)*
    - *Availability -> information has to be consistently accessible (maintaining hardware & infrastructure)*
- *Challenges:*

  - *Despite all this security measurements usability for researchers has to be taken into account*
  - *Ongoing effort of building trust with data providers*
- *New project -> SPHN federated query system (information retrieval system)*
- *The SPHN website provides more information about the contractual framework including a Data Transfer User Agreement*

Dr. Jacques Babel / SFSO, Switzerland

Jacques is head of the Longitudinal Analyses in Education (LABB) study program at the Swiss Federal Statistical Office (SFSO).

- *LABB Study:*
  - *Studies educational trajectories*
  - *Principle of producing & delivering already pre-processed linked longitudinal data & producing meaningful information*
  - *Educational and population register data sources at the core, complemented with peripheral data sources (e.g. labour market data)*
- *Growing demand -> already more than 32 data deliveries in 2022*
- *Too few resources (only ~4.5 FTE) for the large demand*
- *Problems:*
  - *Lack of anticipation -> people don't think about linkage early enough, no centralized storage of identifiers*
  - *Concepts aren't aligned between LABB and other surveys*
  - *Expectations versus real world -> Mismatch between research projects and data request ("we want all data")*
  - *Role of the LABB team in context of research -> no SNFN funding, large investments (data experts)*
- *Potentials:*
  - *Measurements of teacher trajectories*
  - *Projection models based on the data*
- *Solutions:*
  - *LABB is becoming backbone of Swiss education research -> more resources*
  - *More infrastructural efforts have to be made -> it has to be sustainable in the long run*

Discussion

- *Kurt Schmidheiny:*
  - *Regarding personal data -> remote access is coming, it's really important, not only for access, but also for security. I encourage the SFSO to "do it".*
  - *CSAD Solution has the most potential for me, for the Norway microdata approach I see it rather critical.*
- *Georg Lutz:*
  - *Health data is very different from social science, so I advise not to mix that. It should be regarded separately.*
- *Julia Maurer:*
  - *Yes, that is different, but maybe it was grown this way because of the different legal basis. Nevertheless, the sensitivity level of data of a data subject is the same in my opinion.*
- *Georg Lutz:*
  - *We should accept that there is a difference.*
  - *But we can cooperate, especially at the technology level, because it is all very expensive.*
  - *We cannot copy "DataShield" to the social science sector.*
- *Julia Maurer:*
  - *There will be intersections between social and health sciences, we need collaboration. But sure, not all applications will be interchangeable.*
- *Shubham Kapoor:*
  - *The 5 Safes are already a common ground. Besides that, we really need an archive for sensitive data. If data gets deleted after 10 years, for long time archiving there's no initiative. There's no central coordination for that.*
- *Samuel Schütz:*
  - *Why should this be centralized?*
- *Shubham Kapoor:*
  - *In Finland we did this centralized approach too -> it was easier for the researchers. There was one institution for the access and for the preservation.*
- *Pablo Andrés Diaz:*
  - *We have a legal basis for this. The preservation of the present for future generations, that's written in the law.*
- *Julia Maurer:*

  - *I want just to continue…making data available is also very complex. We need agreements from the data provider - in health sector we have strict requirements from the data providers. The data has to stay in the hospitals. For preservation we need technical infrastructures, and we need funding. But also, data providers have to integrate the possibility and the need to reuse the data -> this is very complex and there are many stakeholders who would not agree with a broad reuse of their data.*
- *Ilka Steiner*
  - *We need in Switzerland the "willingness" to share data -> we need real incentives for sharing and providing data -> we need a discussion about the value of data.*
- *Deborah Wiltshire:*
  - *One addition from me to getting the willingness -> a formal archive for such data would mean that the data get a DOI identifier. So that it is searchable. This could be the incentive -> you publicise and get recognition for that (like it is for papers).*
- *Georg Lutz:*
  - *I agree -> those who produce the data are mostly not recognized. So, this has to change -> citing data providers could be a first step.*
- *Sebastian Sigloch:*
  - *What we saw today is that solutions exist for the subject "safe and easier access to sensitive data". The wheel doesn't need to be reinvented, but collaboration is the key -> there needs to be alignment on strategies, governance & standards.*
  - *We need a national infrastructure to preserve the data and governance.*
- *Kurt Schmidheiny:*
  - *We didn't talk about money -> infrastructures are costly. So, the question is "what is the best way of funding?*
  - *Shared solutions are needed.*