FORS

explore.understand.share.

Sensitive data in the Social Sciences and Humanities

Alexandra Stam & Pablo Diaz, FORS data archive

November 7th, 2022

Outline

1. A few words about FORS
2. Processing personal and sensitive data
3. Informed consent
4. Anonymisation
5. Access control

A few words about FORS

Swiss Centre of Expertise in the Social Sciences

FORS
explore.understand.share.

PROJECTS DATA SERVICES TOPICS PUBLICATIONS EVENTS & TRAINING ABOUT FORS

FORS IS THE SWISS CENTRE OF EXPERTISE IN THE SOCIAL SCIENCES.

We produce survey data for national and international surveys.

We provide tools for the information infrastructure in Switzerland and abroad.

We offer consulting services for social science researchers.

We do thematic and methodological research in empirical social sciences.

FIND & DEPOSIT DATA

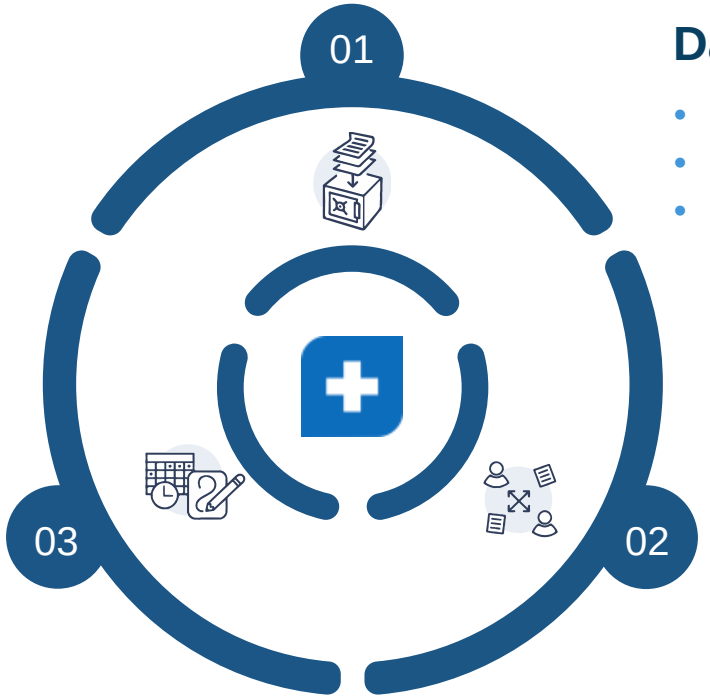
STAFF

OPEN POSITIONS

FORS Data Archive

Data Archiving

- New requirements
- Long-term preservation
- Enhance the value of research projects



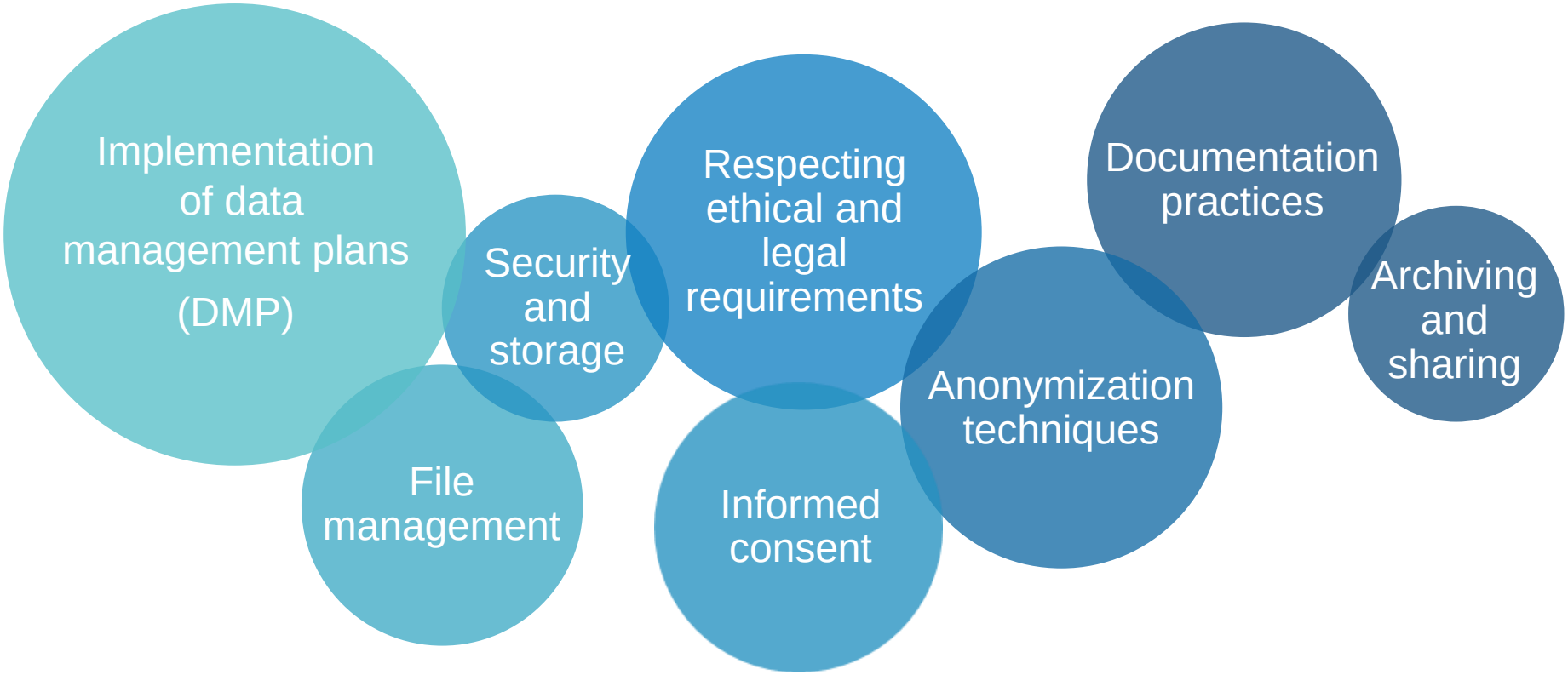
Data Access

- Direct access to datasets and project descriptions

Data Management Support

- Consultancy
- Trainings
- Guides

Data management support



FORS

Processing personal and sensitive data

Processing personal and sensitive data

Scientific research often involves the **processing** of **personal** and/or **sensitive** data

Processing	Personal data	Sensitive data
Any operation with data, irrespective of the means applied and the procedure, and in particular the collection, storage, use, revision, disclosure, archiving or destruction of data. (Art. 3 lit. a FADP)	All information relating to an identified person. (Art. 3 lit. a FADP)	Personal data on: 1. Religious, ideological, political or trade-union; related views or activities 2. Health, the intimate sphere or the racial origin; 3. Social security measures; 4. Administrative or criminal proceedings and sanctions. (Art3. lit. c FADP)

Personal data

"All information relating to an identified or identifiable person" (Art. 3 lit. a LPD)

Very broad notion: everything that can be related to a specific person is personal data !

The most common: name, date of birth, home address, phone number, email, IP address, picture, etc.

But also: opinions, original ideas, quotes, etc.

Personal data can be "objective" or "subjective"

Examples of personal data

Contact details	Quotes	Picture
•First name: Paul •Last name: Dupont •Phone number: 123456 •email: paul@dupont.ch	• « Act as quickly as possible but as slowly as necessary »	

The way you dance is unique, and computers can tell it's you

Nearly everyone responds to music with movement, whether through subtle toe-tapping or an all-out boogie. A recent discovery shows that our dance style is almost always the same, regardless of the type of music, and a computer can identify the dancer with astounding accuracy.

Studying how people move to music is a powerful tool for researchers looking to understand how and why music affects us the way it does. Over the last few years, researchers at the Centre for Interdisciplinary Music Research at the University of Jyväskylä in Finland have used motion capture technology—the same kind used in Hollywood—to learn that your dance moves say a lot about you, such as how extroverted or neurotic you are, what mood you happen to be in, and even how much you empathize with other people.

Even Anonymous Coders Leave Fingerprints

Researchers have repeatedly shown that writing samples, even those in artificial languages, contain a unique fingerprint that's hard to hide.

RESEARCHERS WHO STUDY stylometry—the statistical analysis of linguistic style—have long known that writing is a unique, individualistic process. The vocabulary you select, your syntax, and your grammatical decisions leave behind a signature. Automated tools can now accurately identify the author of a forum post for example, as long as they have adequate training data to work with. But newer research shows that stylometry can also apply to *artificial* language samples, like code. Software developers, it turns out, leave behind a fingerprint as well.

Rachel Greenstadt, an associate professor of computer science at Drexel University, and Aylin Caliskan, Greenstadt's former PhD student and now an assistant professor at George Washington University, have found that code, like other forms of stylistic expression, are not anonymous. At the DefCon hacking conference Friday, the pair will present a number of studies they've conducted using machine learning techniques to de-anonymize the authors of code samples. Their work, some of which was funded by and conducted in collaboration with the United States Army Research Laboratory, could be useful in a plagiarism dispute, for instance, but also has privacy implications, especially for the thousands of developers who contribute open source code to the world.

COMPUTER SCIENCE

Artificial intelligence unmasks anonymous chess players

Software that identifies unique styles poses privacy risks

By **Matthew Hutson**

Think your bishop's opening, queen's gambit, and pawn play are unique? A new artificial intelligence (AI) algorithm has got your chess style pegged.

AI software can already identify people by their voices or handwriting. Now, an AI has shown it can tag people based on their chess-playing behavior, an advance in the field of "stylometrics" that could help computers be better chess teachers or more humanlike in their game play. Alarmingly, the system could also be used to help identify and track people who think their online behavior is anonymous.

"Privacy threats are growing rapidly," says Alexandra Wood, a lawyer at the Berkman Klein Center for Internet & Society at Harvard University. She says studies like this one, when conducted responsibly, are useful because they "shed light on a significant mode of privacy loss."

Chess-playing software, such as Deep Blue and AlphaZero, has long been superhuman. But Ashton Anderson, a computer scientist at the University of Toronto and principal investigator of the new project, says the chess engines play almost an "alien style" That required the system to recognize what was distinctive about each player's style.

The researchers tested the system by seeing how well it distinguished one player from another. They gave the system 100 games from each of about 3000 known players, and 100 fresh games from a mystery player. To make the task harder, they hid the first 15 moves of each game. The system looked for the best match and identified the mystery player 86% of the time, the researchers reported last month at the Conference on Neural Information Processing Systems (NeurIPS). "We didn't quite believe the results," says Reid McIlroy-Young, a student in Anderson's lab and the paper's primary author. A non-AI method was only 28% accurate.

"The work is really cool," says Noam Brown, a research scientist at Meta (the parent company of Facebook) who has developed superhuman poker bots. He looks forward to chess bots that mimic Magnus Carlsen, the reigning world champion, and says style-aware AI could transform other computer interactions. "There's a lot of interest in chatbots, where you can have a chatbot that would speak in the style of Albert Einstein or something," he says.


(DON’T) WORRY YOU ARE
UNIQUE!

Sensitive data

Personal data on religious, ideological, political or trade-union related views or activities; health, the intimate sphere or the racial origin; social security measures; administrative or criminal proceedings and sanctions (Art3. lit. c FADP)

The list provided by the law is exhaustive (e.g. in Switzerland salary is not considered sensitive data)

That said, depending on the context, almost all data can be considered sensitive (name, photo, job, etc.)

Examples of sensitive data

Contact details	Quote	Picture
•First name: Sebastián •Last name: Calfuqueo •Gender: trans •Job: trade unionist	•« Today, our developed nations live in opulence, excess and waste, with the consequences that our environment is degrading and the climate is wreaked » •	

Sensitive data ?

Is public data personal data ?

- Sometimes people make their personal data available to everyone. For example via social networks, blogs, websites, the press, etc.

- Even if this data could be qualified as “public”, it is still personal data and must be treated with care.

- It is important to check the privacy policies of the places where data is collected. Some types of processing may be prohibited.

Managing sensitive data

Due to their special nature, sensitive data need to be managed with extra care.

In order to ensure a sufficient level of protection while maintaining the usefulness and “shareability” of the data, it is possible to combine several measures, the three main ones being:

- Informed consent
- Anonymisation
- Access control

Informed consent

Law and consent

When it comes to collecting and processing **sensitive data**, consent is often the only legal basis available.

- Cantonal research institutions rarely have an explicit legal basis for the collection of sensitive data.
- ETH are among the only research institutions in Switzerland with a clear legal basis.

Law and consent

- Even if there are cases where obtaining consent is not inevitable (processing of non-sensitive personal data) **informating people (participants) is always mandatory** when personal data is collected (directly or indirectly)!

- It is therefore not necessarily easier or lighter to work without consent.

How to ensure that the consent is valid?

Regarding the form of consent:

- Consent may be: **oral** or **written**
  - This said, it's always useful to have proof
  - For HRA research, consent must be written
- Where **sensitive** data is involved, consent must be **explicit**
  - Simply answering a questionnaire, for example, cannot be considered as consent.

What information should be provided?

- The **identity** of the researchers in charge of the project
- **Understandable** statements describing the **purpose** of the research (long-term)
- Contracts with/disclosure to **third parties** (list)
- The right to **access** and **rectify** data
- The guarantee of being **free to decide** not to participate in the project,

What information should be provided II ?

- An **honest and complete description** of the protection/security measures
- The existence of any **conflict of interest**
- **Preservation** and **reuse** of data
- The possibility of being informed of the **results**

Anonymisation

Anonymisation – a definition

- The notion of anonymisation refers to a process by which the elements allowing the identification of a person are definitively deleted from a dataset, a document, an interview transcript, etc.
- Anonymisation represents a principal solution for complying with data protection requirements.
- Legally, this means that an individual cannot be identified without significant effort.

Anonymisation – a difficult promise to keep

- Individuals are more unique than we think!
  *Studies have shown that by **crossing three simple variables**, namely date of birth, postal code and gender, **63% of the US population can be identified** (Golle, 2006).*
- The collection of big data (via apps etc.) makes identification very easy due to the massive nature of unique data collected.
- The ability to cross-reference research data with other datasets, information from social networks, blogs, websites, etc. greatly facilitates (re)identification.

Direct and indirect identifiers

- Direct identifiers alone are sufficient to identify people (e.g., name, AVS number)
- Strong indirect identifiers allow fairly easy identification (e.g., home address, telephone number)
- Weak indirect identifiers allow identification through combinations of variables

Factors to be considered in your strategy

1. The nature and type of personal data to anonymise
2. The future users of the data and conditions of use
3. Balancing utility and data protection
4. Risk management
5. What was promised to respondents

Developing your anonymisation strategy

Your strategy should be developed early in the project and include at least:

- an evaluation of disclosure risk, and
- a description of the anonymisation measures and their rationale.

The strategy will serve as documentation for secondary users. Its implementation should be described after anonymisation has been completed.

Access control

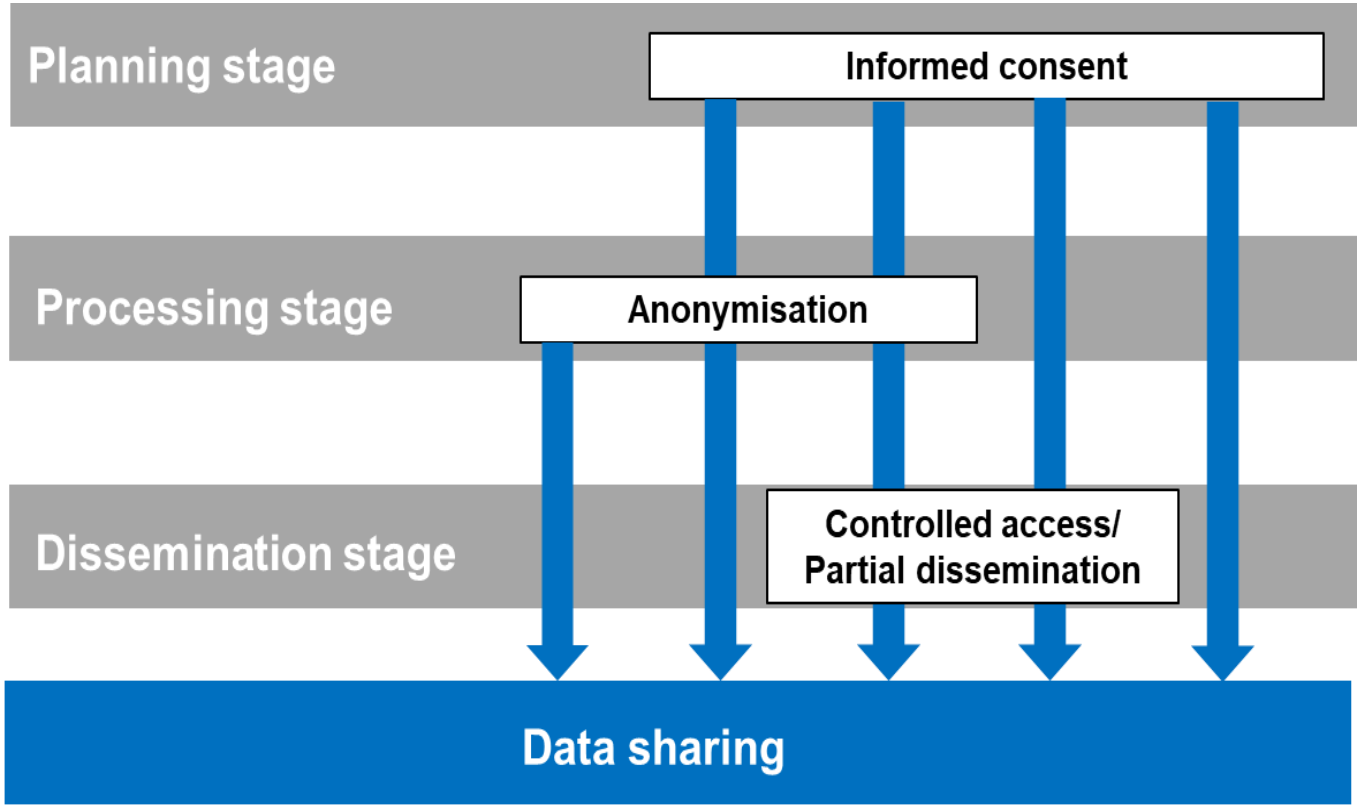
Data access

Not all data can be shared extensively. It is therefore important to define an accessibility policy that is consistent with the risks identified and the measures adopted.

Basically, once archived, the data can be:

- Accessible to anyone
- Conditionally available
- Confidentially protected (embargo)

Data sharing



Ressources

FORS guides

- Ethics in the era of open research data;
- Informed consent;
- How to draft a DMP
- Pre-registration and registered reports
- Data anonymisation, legal, ethical and strategic considerations

https://forscenter.ch/publications/fors-guides/

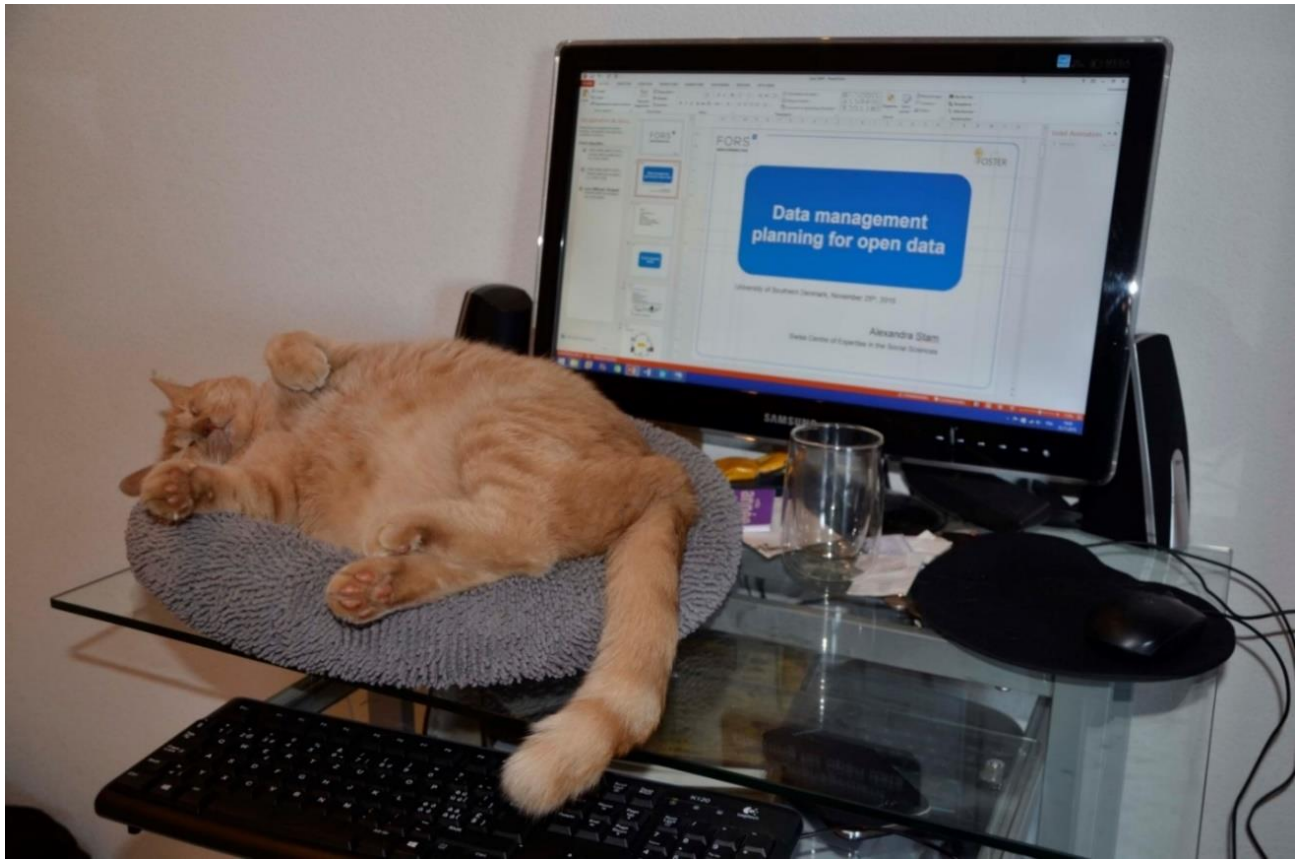
Webinars on data management

https://forscenter.ch/data-management-webinar-series/

Ressources on data sharing data and SWISSUbase:

https://forscenter.ch/data-services/help-resources/

Questions?



Contact: alexandra.stam@fors.unil.ch
pabloandres.diaz@unil.ch

FORS