Lecture notes for Privacy and SIKT for IN5000

Trenton Schulz

14 March 2023

These are notes to the privacy and SIKT proposals for the IN5000 course, qualitative methods.

The point of this lecture is to help students review aspects about privacy that are necessary for doing research in informatics, especially with people. This includes understanding terms about personal data and how to apply for research through the Norwegian Agency for Shared Services in Education and Research (SIKT), formerly NSD. We will do some of this examination by looking at some studies in internet research.

I should emphasize that this is one lecturer's interpretation of the law and regulations. This is not the only interpretations. Some may interpret this more broadly or narrowly. At some point, especially if you continue in research, you will have to come up with your own interpretation as there *will* be disagreements.

One thing that we won't discuss today is how to keep data secure and private. There are mechanisms that one can use such as UiO's Services for sensitive data (TSD). If you are connected to a project that is using TSD, you'll likely get your own documentation about it. There are also simpler solutions that you can consider (such as storing on an encrypted hard drive), but there are other things one needs to keep in mind (such as securing the network). I will only encourage that if you have access to TSD, please try to use it. I'll also mention that information gets old as time. I wrote a guide about keeping data secure based on best practice of the time (2011), that method is now considered insecure.

UiO also has documentation that describes how to classify data and possibilities for storage that may be useful for answering some questions around this.

1. LAWS & GDPR

The Norwegian privacy rules have a basis several laws. The Nuremberg Code, which was created after World War II was the first work on informed consent of a participant in research. The Deceleration of Helsinki from the 1960s, which set up a set of ethical principles for human experimentation. There is also the Belmont Report from 1978 which establishes ethical principles for human subjects in biomedical research. Unfortunately, this is was an attempt to make a one size fits all solution that really doesn't work for all kinds of research. There is also the UN Declaration of Human Rights and European Convention on Human Rights that help form the basis of the guidelines.

Today most of discussion around personal data is related to the General Data Protection Regulation (GDPR). The regulation took effect in June 2018. Many of the privacy regulations are the same as before, but its even more specified for personal information. It includes such things as rights such as the right to be forgotten, that privacy is "built in", not something that can be added. GDPR defines how to document that informed consent has

been given and what needs to be included. There are also routines with warning about data breaches and it defines strong punishments for not following the regulations.

Norway implements GDPR through the personal data law *personapplysningsloven* (LOV-2018-06-15-38). There is also a translation of the GDPR into Norwegian that starts in chapter 9.

We experience consequences of the GDPR every day when we go online and use a website (or a mobile app). Websites like DNB are required by the GDPR to explain how they use information they collect about you (cookies). An interesting examination that can be done is how misleading most of these interfaces are to encourage you to give them all your information, but that is another story.

GDPR has several terms that are good to know when working with data about people:

Data subject

(*den registrerte*) is a natural person.

Personal data

(*personopplysninger*) is any information relating to an identified or identifiable natural person (either directly or indirectly).

Processing

(*behandling*) is any series of operations done to personal data.

Controller

(*behandlingsansvarlig*) is the person who determines the purpose of processing of the personal data and the means of processing.

Processor

(*databehandler*) is the person who processes the personal data on behalf of the controller. This can also be a cloud service provider (e.g. TSD).

So, if you (the controller) are going to be collecting and processing personal data from data subjects, the GDPR has specific requirements for the controller. That is, if you are recording or processing of information about people by electronic means or creating a manual register containing special categories of personal data. You, as a researcher, must fill out a special form to the privacy ombudsman for research (*personvernombudet for forskning*).

Discussion point: Think about your research projects. Do you think that you will be collecting personal data? Why or why not? Are there ways that you think can avoid collecting personal data? Is it "bad" to collect personal data for a project?

2. THE NOTIFICATION FORM

Filling out the notification form (*meldskjema*) is straightforward. There are just a couple things to keep in mind.

Plan your study carefully

roundtrips between you and SIKT will take time

Have your paperwork in order

all documents for your activity need to be included

A standard evaluation normally in done within 30 days

no data collection can be done until the form is processed.

When in doubt talk with your advisor or contact the SIKT

Your advisor and SIKT are there to help you comply with the law. They are not there to stop your project or play gotcha.

Vike & L'organge Fürst (2020) put forth an argument that the notification forms from SIKT "reduces" questions of privacy and ethics to only being about personal data and neglects other aspects of ethics, which I tend to agree with. They also argue that the SIKT requirement for preparation goes against what athropologists typically do, and that anthropologists are often "on location" to observe and don't have research questions formed at the start. They often are exposed to extra information about people that is collected, but not really processed like it would be for medical studies.

What happens if you need to make a change? Well, you may not have to. If you are doing interviews on a topic and you are adding or removing questions that aren't collecting a new type of data, you don't need to send in a new form. If you are suddenly collecting new types of data (e.g., you didn't say you wanted to video record, but now you do), then you need to notify SIKT about the change. This is normally just a miniature form of the previous form where you describe the change and provide a new informed consent form. The changes may still require a standard evaluation taking up to 30 days, but it usually is a little shorter.

What about the *Regional komiteer for medisnsk og helsefaglig forskningsetikk*) or REK? In general, if you are *not* collecting health data, you probably *do not* need approval from REK. Regardless, REK's approval only looks at the health and medical aspects of the research, it does not look at how data is stored. So, you will still need to fill out a notification form for the SIKT for processing personal data. When in doubt, discuss with your advisor.

2.1. Types of evaluation

When sending in the notification form, there are three types of notifications that can happen.

Standard evaluation

This means that an advisor at Sikt goes through the form and evaluates it. This should be completed in 30 days or less, but normally is around 30 days. If the advisor believes that the processing of personal data has major privacy issues and risks, the person may have to fill out a privacy consequences evaluation (this is normally seldom for master projects).

Automatic evaluation

The form is evaluated by a machine with no advisor looking at the form. This is done for forms that collect personal information that has low risk of causing issues. There are multiple criteria, but the big points are:

- The project does not collect special categories of personal data.
- The length of the project is limited
- The number of participants is low
- All participants get individual information
- All participants are over 15-years old

The attachments to the forms (e.g., interview guides and informed consent forms) are not examined. Students and researchers should use templates to ensure that all legal information is present. Although it is automatic, there are still random inspections of a form to see how well the automatic evaluation is working correctly. The most common problems are that the informed consent forms are perhaps incorrect, that special categories of personal information is collected, or that personal information about third-party people is collected.

The service started in November 2022 and it is likely that many master projects would qualify for automatic evaluations. The evaluation is completed in one day!

Group evaluation of student project

I mention this for completeness, but this is more for bachelor courses or a year-long course where students will be running studies that are short and have similar topics. Then, the teacher of the course can consider submitting a notification form for a whole class. This is *not* relevant for master projects.

Sikt receives over 11,000 projects from 130 research and education intuitions to evaluate each year, so they hope that they can reduce time with the automatic evaluation. Since 2023 is the first year with real automatic evaluation, we'll see how that goes.

3. INTERNET RESEARCH ETHICS

The four issues that we looked examined in the lecture were:

1. Is online interpersonal media (social media) considered public or private information?
2. Conducting covert research vs. informed consent
3. Protecting anonymity in the face of the Mosaic Effect
4. Handling the raw data from internet research.

For issues 1 and 2, we discussed with Jernigan & Mistree's Gaydar study (2009), which used "public" information on Facebook to help expose a person's sexual orientation. Their research showed how public information could be linked together, but the people did not give their consent.

The Association of Internet Research suggests that informed consent is not necessary when data is collected from the public sphere with no intervention from the persons whose activities are observed and recorded, **and** the collection data does not include personal

identifiers, which if released, could result in reputational or financial harm to the person whose activities are observed.

We also discussed some issues in protecting anonymity. While this is straight forward for one data set, it becomes more difficult when more data sets are included. The more different data sets can be linked together, the easier it is to deanonymize the data. This is called the Mosaic Effect. We looked at the examples of AOL's "anatomized" search data, and an example of how a master student's attempt at anonymizing data fails with later technology.

While anonymization is often talked about with structured data in tables that may be relatively straightforward to anonymize. Weitzenboeck and colleagues highlight how it is even more difficult for unstructured data as text or images cannot be easily scanned and redacted. Weitzenboeck and colleagues also discuss different definitions of what constitutes anonymized data.

Weitzenboeck and colleagues (2022) also look at two methods for determining how identifiable the data is: a risk-based approach and a strict approach based on the Working Party's Opinion on Anonymization Techniques. The risk-based approach uses the idea of a *motivated intruder*, a reasonable competent person with access to libraries and internet, but not advanced hacking knowledge, to determine how likely it is that the data will be collected by someone outside and identified. Given that the motivated intruder does not have unlimited budget and skills, one can come up with a reasonable idea of the risk involved with storing and anonymizing the data. It also works with the idea of reasonable risk for the technology available at the time. The Working Party's Opinion is very strict and basically says that anonymous data cannot exist if the original data exists anywhere. Linking back to the original with text is even easier. This basically means that all original data needs to be destroyed or the data is presented completely anonymous and useless.

Finally, for issue 4, having the raw data available is useful so that results can be verified. To assist in peer-review and a possibility for helping in replication a study, raw data should be available on request. Keep the data, but pseudonymize the records using different numbers of real IDs. Keep raw data access restricted.

You can read *Ethnic Guidelines for Internet Research* for more information of about internet research.

REFERENCES

Jernigan, C., & Mistree, B. F. T. (2009). Gaydar: Facebook friendships expose sexual orientation. *First Monday*, *14*(10). https://doi.org/10.5210/fm.v14i10.2611

Vike, H., & L'orange Fürst, E. (2020). Forskningsetikk og forskningens frihet: Utfordringerfor antropologifaget. *Norsk antropologisk tidsskrift*, *31*(3), 165–176. https://doi.org/10.18261/issn.1504-2898-2020-03-02

Weitzenboeck, E. M., Lison, P., Cyndecka, M., & Langford, M. (2022). The GDPR and unstructured data: Is anonymization possible? *International Data Privacy Law*, ipac008. https://doi.org/10.1093/idpl/ipac008