Internet research (definition) + Processing personal data and the law + Ethics when doing Internet research

Gisle Hannemyr
INF5220, spring 2015

Overview of lecture

- Definition of Internet research (i.e. collecting field data from the Internet).
- Examples of Internet research.
- Processing personal data: Legal requirements in Norway.
- Processing personal data: Ethics

"Internet research" roughly encompasses inquiries that:

1. Study how people use and access the Internet, e.g., through observing ongoing activities or participating on "social" network sites, listservers, web sites, blogs, games, virtual worlds, and other online environments or contexts.
2. Utilizes the Internet to collect data from and about users, e.g., through online interviews, surveys, archive analysis, or by scraping or other automated means.
3. Access and analyse already collected data in the form of databanks, curated collections and/or other data repositories available via the Internet.
4. Employs discourse analysis, semiotic analysis, visual analysis, or other methods of content analysis to study information that exists on the web and/or internet-facilitated images, writings, and other online media forms.
5. Studies large scale production, use, and regulation of the internet by governments, industries, corporations, and military forces.
6. Engages in web software engineering to study software, formats, standards, and other relevant technologies, including their properties and use.
7. Examines the design and/or infrastructure of systems, interfaces, web pages, design elements, technical components and assemblages.

Internet research and personal data

- Research into the social, cultural, political, economic, ethical, technical and aesthetic aspects of the Internet often involves interactions with, and/or collections of *personal data* from and about *individual persons* from online sources, or from others sources, for data processing purposes.
- Basics: Same legal/ethical framework as any "other" research involving processing of personal data.
- However, some special problems.

Examples of Internet research

- Discourse analysis of online archives, e.g. blogs, online newspapers (including comment fields), online "social" media, etc.
- Ethnographic research into virtual "communities".
- Using robots to scrape and analyze online data from individuals posting online.

Example: Online archive analysis

- Eric Monteiro: *Scaling information infrastructure: the case of the next generation IP in Internet.* The Information Society, 14(3):229-245, 1998
  - A case study of the development of IP ver. 6.
  - Based (mostly) on analysing the email archives available online that the design board left behind.

Example: Discourse analysis

- Maja van der Velden and Alma Culén: *Information Visibility in Public Transportation Smart Card Ticket Systems.* (2013)

  Paper contrasting the visibility of ticket information to users of paper tickets and smart card tickets, based upon comments scraped from social media.

  ”From the introduction of the smart card ticket (Reisekort) in 2009 until September 2013, 392 articles were published [13]. These articles usually triggered many comments from readers.”

Example: Ethnographic analysis of V.C.

- Christine Hine: *Virtual Ethnography;* Sage 2000
  - This is an anthropological study centred on a single event: the 1997 US trial of British nanny, Louise Woodward. It focuses on the role of the Internet, concentrating particularly on web sites and newsgroups that were created and used in the frenzy of media interest that accompanied the trial. Its discussion of space and time, identity and authenticity set up some intriguing discussions about prevailing attitudes among Internet users and how the Net functions both as a cultural tool and as a micro-culture in itself.
  - The book also discusses methods and practices of ethnographic research on the Internet.

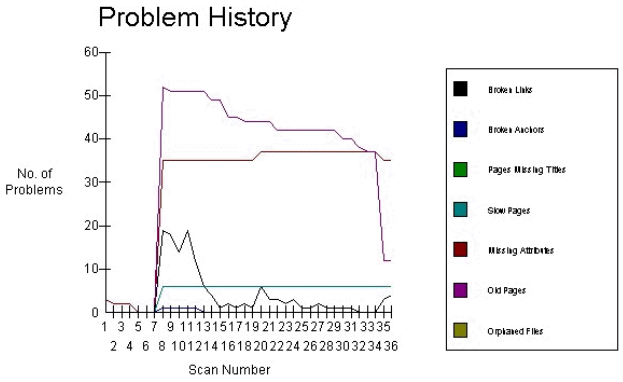
Semi-automatic semantic analysis using robots

- Carsten Sørensen et al anlysis of Apple iStore acceptance and rejectance decision.
- Based upon using a web robot to capture a large number of "war stories" about publishers dealing with Apple, and using semi-automatic analysis of this data to extract the unpublished "rule set" used by iStore.

Robot analysis

- Framework for managing a "collection" of robots and tools for analysis to measure certain aspects of the World Wide Web:
- Will accumulate data along the following axes:
  - Page size and page complexity and content (media, links, etc.)
  - Size, growth, rate of change
  - Problems: (broken links, etc.)
  - Quality (Latency, Packet loss, Reachability)
  - Adoption of the «Semantic Web»-vision
- Background:
  - Bharat, K. and Broder, A. (1998) A technique for measuring the relative size and overlap of public Web search engines, In: *7th International World-Wide Web Conference,*Elsevier Science, Brisbane, Australia, 14-18 April.
  - Lawrence, S. and Giles, C. L. (1999) Accessibility of information on the web, *Nature,* vol. 400, pp. 107-109.

Example: Robot analysis (software engineering, not “field work”)



Example: Robot analysis

- Warren Sack: Discourse Architecture and Very Large-scale Conversation; in: Sassen and Latham (eds.): *The Digital Order,* Princeton University Press 2005
  - Sack introduces DA and VLSC as concepts. He then uses robot analysis of available online settings (e.g. Usenet newsgroups) to “map” conversations into semantic networks (to identify key themes), and into conversation clusters (to identify social networks).
  - He is aware of the ethical problems posed by analyzing on line conversations among individuals about sensitive topics such as politics. His solution is to make make sure his tool only show very high stylized graphics of themes and social networks, so that no personal data is exposed.

Regulatory framework for Internet research

- Because Internet research may involve use of personal data, there may be legal and ethical guidelines that regulates IR data processing:
  - EU privacy directive
  - Norwegian personal data act (poppl.)
  - European Convention on Human Rights
  - UN Declaration of Human Rights
  - The Nuremberg Code
  - The Belmont Report
  - The Declaration of Helsinki

Legal requirements in Norway

- The legal requirements for the controller doing research where *personal data* are collected and processed are specified in *Personopplysningsloven* (oppl.):
  - Main requirement: *All* such research need to be reported on a special form to *Personvernombudet for forskning* (Privacy ombudsman for research).
  - http://www.nsd.uib.no/personvern/en/notification_duty/
  - http://www.nsd.uib.no/personvern/en/notification_duty/test
- My guidelines about filing a report (in Norwegian):
  - http://heim.ifi.uio.no/~gisle/ifi/pol.html

Norwegian personal data act §2: Definitions

- **personal data** (*personopplysning*): any information and assessments that may be linked to a natural person,
- **processing of personal data** (*behandling av personopplysninger*): any use of personal data, such as collection, recording, alignment, storage and disclosure or a combination of such uses,
- **personal data filing system** (*personregister*): filing systems, records, etc. where personal data is systematically stored so that information concerning a natural person may be retrieved.
- **consent** (*samtykke*): any freely given, specific and informed declaration by the data subject to the effect that he or she agrees to the processing of personal data relating to him or her.

Actors in Norwegian legal framework (POL)

- **controller** (*behandlingsansvarlig*): the person who determines the purpose of the processing of personal data and which means are to be used (POL §2.4).
- **processor** (*databehandler*): the person who processes personal data on behalf of the controller (POL § 2.5).
- **Data subject** (*registrerte*): the person to whom personal data may be linked (POL §2.6).

Consent must be:

- Freely given:
  - No pressure or coercion or linking to favours
- Specific:
  - Usually by signature
- Informed:
  - Purpose of reserach
  - How personal data will be used
  - When personal data will be destroyed or anonymized.

Poppl. § 8: Conditions for the processing of personal data

Personal data (cf. section 2, no. 1) may only be processed if the data subject has consented thereto, or there is statutory authority for such processing, or the processing is necessary in order:

a) to fulfil a contract to which the data subject is party, or to take steps at the request of the data subject prior to entering into such a contract,
b) to enable the controller to fulfil a legal obligation,
c) to protect the vital interests of the data subject,
d) to perform a task in the public interest,
e) to exercise official authority, or
f) to enable the controller or third parties to whom the data are disclosed to protect a legitimate interest, except where such interest is overridden by the interests of the data subject.

«Personopplysning» = Personal data

- POL: Data that may *directly or indirectly* connected to a physical persion
  - Name
  - PIN
  - IP-address
  - Patient profile of a rare disease + location (mosaic effect - *bakveisidentifisering*)

The Mosaic Effect

"The Mosaic Effect occurs when the information in an individual dataset, in isolation, may not pose a risk of identifying an individual (or threatening some other important interest such as security), but when combined with other available information, could pose such risk. Before disclosing potential PII [personally identifiable information] or other potentially sensitive information, agencies must consider other publicly available data – in any medium and from any source – to determine whether some combination of existing data and the data intended to be publicly released could allow for the identification of an individual or pose another security concern."
Source: http://project-open-data.github.io/policy-memo/

POL: Report form compulsory if:

- Recording or processing of information about individuals by *electronic* means.
  - NB: "electronic" ⇔ "digital". Analogue recording is not considered "electronic" for legal purposes.

- or -

- A manual register containing *sensitive personal data* will be created.

POL: Permit compulsory if:

- Sensitive personal data is recorded (POL § 33).
- Sensitive personal data (POL §2.8) are data that reveals information relating to:
  - racial or ethnic origin, or political opinions, philosophical or religious beliefs;
  - the fact that a person has been suspected of, charged with, indicted for, or convicted of, a criminal act;
  - health;
  - sex life;
  - trade-union membership.

POL: But permit not compulsory for research if the *privacy ombudsman for research* approves:

1. First time contact to selection of respondents is based upon, either:
   - publicly available data (i.e. data that exists in the public sphere);
   - a responsible person at the insitution where the respondent is registered;
   - initiative from the respondent.
2. The responent has given *valid consent* to all parts of the research.
3. The project is terminated at the time agreed upon.
4. All material collected is destroyed or anonymized when the project is terminated.
5. The project is not joining data from external registers or data bases.

POL: Reuse of data is *not* permitted without new consent

- POL § 11c. The controller shall ensure that personal data which are processed … are not used subsequently for purposes that are incompatible with the original purpose of the collection, without the consent of the data subject.

Valid consent

- **Informed** – data subject must understand what consent implies (what data is collected and how they shall be processed), and what the consequences of consent is for the data subject.
- **Voluntary** – data subject must not be punished or have rights taken away if consent is not granted.
- **Explicit** – data subject must perform some affirmative act to express consent.

Internet data collection

- Special challenges
  - Method
    - How to locate, select, verify and document data.
  - Ethics
    - Conducting research enframed in a set of sound ethical guidelines.

Summary (from AOIR) of difficulties in Internet Research (IR)

1. Difficulty of ascertaining subjects' identity because of use of pseudonyms, identity-games, etc.
2. Difficulty in obtaining informed consent from online subjects.
3. Difficulty in discerning correct approaches because of a greater diversity of research venues (email, chat rooms, web pages, etc.)
4. Difficulty of discerning correct approaches because of the global reach of CMC (engaging people from multiple cultural settings).
5. Difficulties posed by covert research (observing subjects that do not know that their behaviours and communications are being observed and recorded) – simply because of the easy access there is to online material ready to capture.

Person or persona?

- In many online environments (e.g. "home" pages, real and faked web media pages, discussion forums, chat rooms, MUDs and MOOs), expression of identity (including multiple selves, avatars and other forms of intentional identity-games) is often constituted through the construction and reception of texts and (sometimes) imagery.
- To a researcher, what is identity in such contexts? Do we need to separate between the "real" (whatever that is) person and the projected "online" persona?

"On the Internet, nobody knows you're a dog."

Public and private sphere: Jürgen Habermas (1992)

- The public sphere: The sphere where matters of public interest is discussed.
  - The arena of mass media is the public sphere.
- The private life: The sphere for everything that is not of public interest, and that does not need to be subject to public inquiry..
  - Diaries, private letters, family photographs and private sound recordings.

Public and private sphere: Do they converge?

- From latin: *con vergere* = «to run together, to incline».
- Concept closely linked to the ongoing digitisation of various types of information- and communication technology.
  - Streams of information that previously has existed in different domains (private letters, newspapers, photographs, vinyl records, telephony, television, etc.) are all entering the digital domain (i.e. based upon digital encoding of information).
- This means that we are distributing a number of *dissimilar* services and information streams that previously has been separate over the *same* digital communication networks.

Divergence

- The "new" media are often far more complex and diverse than the "traditional" media, both in terms of genre, manufacturing, distribution of responsibilities, roles and identity.
- This diversity is an enrichment, but it also creates challenges in relation to privacy and data protection.

Mass media defined by McQuail (2000, p. 4)

The term "mass media" is shorthand to describe means of communication that operate on a large scale, reaching and involving virtually everyone in society to a greater or lesser degree. It refers to a number of media that are now long-established and familiar, such as newspapers, magazines, film, radio, television and the phonograph (recorded music). It has an uncertain frontier with a number of new kinds of media that differ mainly in being more individual, diversified and interactive and of which the Internet is the leading example.

Interpersonal media (aka. social media)

- Privacy ombudsman for research insists data data collected from interpersonal media (social media) belongs in the private sphere.
- PFU thinks data collected from interpersonal media belongs in the public sphere if it is visible to a large number of persons.

Facebook i bladet «Tromsø»


Det er reservert et sted innerst i helvete for israelittisk lederskap

PRESTE-STUDENT: Eirik Junge Eliassen.

Kjenner du oss - kjenner du byen

iTromsø

Side 5

NR. 122 TIRSDAG 1. JUNI 2010 UKE 22 11. ÅRGANG 20,- ØKER MEST I NORD-NORGE: 42.000 DAGLIGE LESERE WWW.ITROMSO.NO

I kjennelsen (PFU-sak 117/10) heter det:

Klagen gjelder et oppslag *i iTromsø* der et innlegg på en lukket profil på Facebook ble gjengitt. … iTromsø anfører at det er snakk om en kjent person i Tromsø og at han har 2.700 venner på sin lukkede profil.

I det foreliggende tilfellet er det snakk om et lukket område der klageren i prinsippet har anledning til å bestemme hvem som skal ha adgang. Det må i denne sammenheng være naturlig å håndtere private ytringer fremmet i en liten lukket gruppe med noen få medlemmer annerledes enn ytringer av allmenn interesse som framkommer i en lukket gruppe med 2500 «venner».

iTromsø var i sin fulle rett til å viderebringe klagerens politiske utsagn om staten Israels ledelse. Ytringen hadde aktuell interesse, og klageren fikk selv anledning til å kommentere utsagnet. Slik utvalget ser det, må han tåle at iTromsø viderebrakte hans synspunkter, selv om han selv fjernet dem fra sitt nettområde etter kort tid.

"Public" according to the General Civil Penal Code (Straffeloven)

- The general civil code was changed on May 24 2013, and now contains a definition of public acts in the 2nd part of § 7:
  - "An act is considered to be committed in public when it is committed by publication of printed matter or in the presence of a large number of persons or under such circumstances that it could easily have been observed from a public place and is observed by any person present there or close to it."
- The preparatory work (Ot.prp. 90 (2003-2004) sec. 12.2.2) refers to case law, and says "a large number of persons" can be interpreted as "about 20-30 persons".
- **NOTE:** The civil penal code describes punishable public acts (such as libel, and hate speech). What is considered "public" acts according to the civil penal code is not necessarely *also* considered "public" data in the context of privacy law or research.

IR ethics, sources:

- Cheltenham and Gloucester College of Higher Education: *Research Ethics: A Handbook of Principles and Procedures.*
- Association of Internet Researchers (AoIR), reports on *Ethical and Legal Aspects of Research on the Internet*
  http://aoir.org/reports/ethics.pdf (2002)
  http://aoir.org/reports/ethics2.pdf (2012)
  - Adapted from biomedical research.

Four major problems

- Is online interpersonal media (social media) in the Public or Private Sphere?
- Covert research/Informed consent
- Protecting anonymity
- Raw data

Covert research methods

- Online research poses in general a risk to individual privacy and confidentiality because of greater accessibility of information about individuals, groups, and their communications – in ways that would prevent subjects from knowing that their behaviours and communications are being observed and recorded (e.g.: a large-scale analysis of postings and exchanges in a USENET newsgroup archive, in a chat room, etc.).

Valid consent is a required for sharing personal data

[P]rivacy is considered widely as a crucial norm in ethical research […] Data arising from research should ordinarily be considered confidential and may not be shared with others without the consent of the researched.

– *Research Ethics Handbook*

Protecting anonymity

[R]esearchers must take care where the alteration of contexts may reveal the identity of data sets hitherto protected. Particular care should be taken with data that arises from covert […] research methods […].

– *Research Ethics Handbook*

Protecting raw data

- Good research practice means that the raw data (for aggregated, pesudonymized or anonymized data that is published) must be available for scrutiny and peer review upon request.
- Solution: Retain the raw data, but pseudonymize records by using numbers instead of real IDs. Make access to RAW data very restricted (encrypted and in custody of a trusted third party, analogous to safekeeping sensitive data accumulated in biomedical research).

Institutional setting

- In biomedical resarch, the institutional setting (i.e. the research clinic) usually has well developed procedures and mechanisms for handling, anonymizing and protecting personal data originating from research.
  - This is taken as given both by the researchers and also by the data subjects (i.e. the patients).
- In Internet research, no similar setting usually exists and has to be constructed by the data controller as part of his/her research framework for each project.

AOIR suggestion:

- Researchers need not obtain informed consent, etc., from subjects if:
  - The data is collected from the public sphere with *no intervention* with the persons whose activities are observed and recorded.
  - The *collection of data* does not include personal identifiers which, if released, could result in reputational or financial harm to the person whose activities are observed.

Handling ethics: MIT "Gaydar" project

"Our analysis demonstrates a method of classifying sexual orientation of individuals on Facebook, regardless of whether they chose to disclose that information. Facebook users who did not disclose their sexual orientation in their profiles would presumably consider the present research an invasion of privacy. Yet this research uses nothing more than information already publicly provided on Facebook; no interaction with subjects was required. Although we based our research solely on public information, only a limited subset of our results, which contain no personally identifiable information, is presented in this paper to maintain subject confidentiality."

Source: Carer Jernigan and Behram F.T. Mistree: *Gaydar: Facebook Friendships Expose Sexual Orientation;* First Monday 14:10; 2009.

Data collected only from the public sphere, but disclosure of personal identifiers could lead to harm for data subjects. The researchers treated their data anonymously, never using real names except to validate their predictions during data analysis. The only copy of the raw data was on an encrypted DVD that was held by their advisor. The project was reviewed ethical review board at MIT and approved.

Why is Internet research so special?

Example: Handling ethics

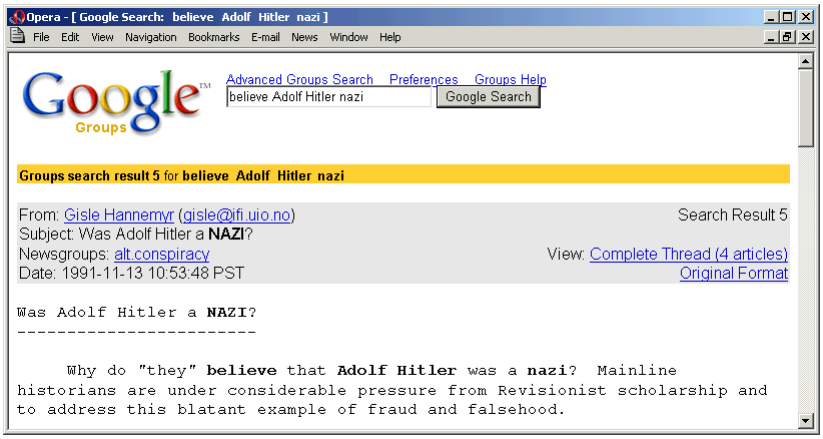
Espen Munch: *En antropologisk analyse av elektronisk nettkommunikasjon,* master thesis in social anthropology at UiO, 1997:

"[Jeg har] valgt å anonymisere både deltakere og grupper i den grad det er mulig i denne oppgaven. Jeg har laget fiktive navn til gruppene, og tatt bort de riktige navnene til opphavsmennene for siterte postinger. Istedenfor ekte aktørnavn har jeg brukt psevdonymer med fiktive fornavn. For at postingene ikke skal bli for lette å spore i News-arkiver, har jeg også fjernet de nøyaktige postingstidspunktene, alt som har med avsenderens epostadresse å gjøre, og eventuelle artikkelnummer."

Pseudonymizing a direct quote

```
From: [John Doe]
Subject: Was Adolf Hitler a NAZI
Newsgroups: [some.newsgroup]
Date: [withheld]
Was Adolf Hitler a NAZI
----------------------
Why do 'they' believe that Adolf Hitler was a
nazi?  Mainline historians are under considerable
pressure from Revisionist scholarship and to
address this blatant example of fraud and
falsehood.
```

... but not very succesfully



Note: Google Groups no longer reveals email addess.

Final words

- The greater the vulnerability of the data subject, the greater the moral obligation of the researcher to protect the data subject from harm.
- Because "harm" is defined contextually, ethical principles are more likely to be understood inductively. That is, rather than universal predicates, doing ethical Internet research requires practical judgment paying attention to context (what in Aristotelian ethics is identified as φρόνησις – *phronēsis* - or practical wisdom).
- When making ethical decisions, researchers must balance the privacy rights of the data subjects with the social benefits of the research and researchers' rights to conduct research. In different contexts the privacy rights of subjects may outweigh the benefits of research.