

# Notes on Internet Research

---

Gisle Hannemyr  
INF5220, Oktober 2007

## Overview of lecture

---

- Legal requirements in Norway
- What *is* Internet research?
- Online resources and bibliographies
- Internet field work
- Ethical issues

## Legal requirements in Norway

---

- The legal requirements for doing research that where personal data about individuals are collected and processed are specified in *Personopplysningsloven* (POL):
  - Main requirement: *All* such research need to be reported on a special form to *Personvernombudet for forskning* (Privacy ombudsman for research).
- Report form guidelines (in Norwegian):
  - <http://heim.ifi.uio.no/~gisle/ifi/pol.html>

## POL: Report form compulsory if:

---

- Recording or processing of information about individuals by *electronic* means.
  - NB: "electronic"  $\Leftrightarrow$  "digital". Analogue recording is not consider "electronic" for legal purposes.
- *or* -
- A *manual* archive containing sensitive personal data will be created.

## POL: Permit compulsory if:

---

- Sensitive personal data is recorded.
- Sensitive personal data is data that reveals:
  - Racial or ethnic background
  - Political, philosophical or religious opinion
  - Criminal record
  - Health related information
  - Sexual relations
  - Membership to trade unions

## POL: But permit not compulsory if:

---

1. First time contact to selection of respondents is based upon, either:
  - publicly available data;
  - a responsible person at the institution where the respondent is registered;
  - initiative from the respondent.
2. The respondent has given informed 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 more than one register or data base.

## Internet Research: An overloaded term

---

- The term "Internet Research" or "Online Research" appear in the relevant literature to denote in a number of *different* contexts and research approaches, such as:
  - Resource discovery
  - Form-based data collection
  - Research *about* Internet (usage)
  - "Field" work (The field=Internet)

## Resource discovery

---

- Using the Internet to search for written documents (e.g. books, articles or manuscripts). The purpose of this search is to locate the written documents and then obtain physical copies through interlibrary loan or similar means.
- Locating various types of information resources that exists the web. This type of search locates documents that you can view on your computer screen.

## Resource discovery

**Book:** Niall Ó Dochartaigh: *The Internet Research Handbook*; Sage 2002.

Table of Contents:

- Research Tools
- Searching for Books and Articles
- Making Contact
- The Web
- Searching by Subject
- Searching the Keyword Search Engines
- Classification, Evaluation and Citation
- Patricia Sleeman (National Digital Archive of Datasets in London): Archives and Statistics

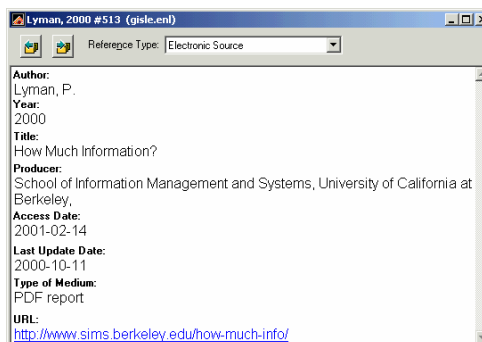
Okt. 2007

INF5220

Page #9

## Online resources Maintaining bibliographies

- BibTex/EndNote
- Keep all your bibliographic references in a database.
- Learn how to change output style (Harvard, IEEE, etc.)



Okt. 2007

INF5220

Page #10

## Online resources

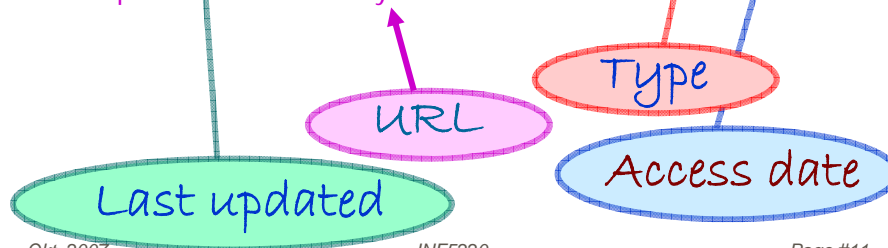
### Citing electronic resources

---

Hannemyr, G. (2002) *Re: Citing electronic sources in scientific papers*, [2002-12-06], (private email message).

Hannemyr, G. (2003) *NCC incident*, [2003-05-13], (email interview).

Lyman, P., et al. (2000) *How Much Information? 2000*, [2001-02-14], last updated: 2000-10-11, School of Information Management and Systems, University of California at Berkeley, (PDF report) <<http://www.sims.berkeley.edu/how-much-info/>>.



Okt. 2007

INF5220

Page #11

## Form-based data collection

---

- Collection of survey-type data through the Internet by the means of web-forms.
  - **Book:** C. Mann & F. Stewart: *Internet communication and qualitative research. A handbook for researching online*; Sage 2000.
    - ◆ Covers mainly issues concerning conducting online interviews and using online surveys.

Okt. 2007

INF5220

Page #12

## Research *about* the Internet

---

- Traditional sociologic or ethnographic study focusing on a particular group or society's use of the Internet.
  - **Book:** Daniel Miller and Don Slater: *The Internet: An Ethnographic Approach*; Berg 2000
    - ◆ In depth regional study (Trinidad) of Internet usage, culture and consumption.

## Field work

---

- Doing "field work"-type data collection on the Internet.
  - **Book:** Steven G. Jones (Ed.): *Doing Internet Research. Critical Issues and Methods for Examining the Net*; Sage 1999
    - ◆ Compilation of 13 articles + Introduction

## Outline – rest of this talk

---

- We're done with:
  - Resource discovery
  - Form-based data collection
  - Studies of Internet usage and consumption
- To be continued:
  - Internet field work
    - ◆ Examples
    - ◆ Ethical issues

## Examples of Internet field work

---

- Analyzing online archives
- Conversations on boards and chat-channels
- Ethnographic research into virtual communities
- Analyzing Internet pages as media expressions
- Using robots to collect and analyze online data (also quantitative)



## Example: 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 analyzing the archives available online that the design board left behind.

## Example: Ethnographic chat analysis

---

- Nancy K. Baym: *Tune In, Log On. Soaps, Fandom, and Online Community*, Sage, 2000
  - An ethnographic study of an Internet soap opera fan group
  - Bridging the fields of computer-mediated communication and audience studies, the book shows how verbal and nonverbal communicative practices create collaborative interpretations and criticism, group humour, interpersonal relationships, group norms, and individual identity.
  - While much has been written about problems and inequities women have encountered online, Baym's analysis of a female-dominated group in which female communication styles prevail demonstrates that women can build successful online communities while still welcoming male participants.

# Example: Virtual communities

- 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.

Okt. 2007

INF5220

Page #19

Media expressions

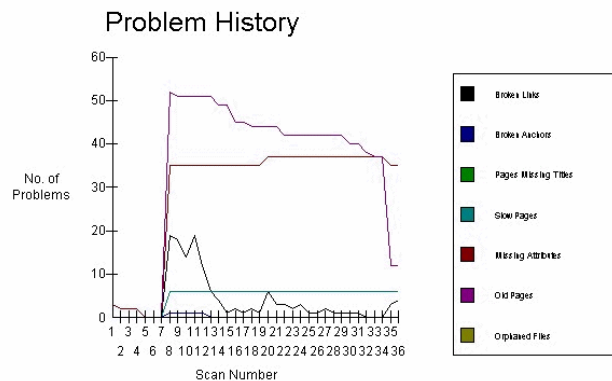
Okt. 2007

INF5220

Page #20

## Example: Robot analysis

---



Okt. 2007

INF5220

Page #21

## Robot analysis (master thesis)

---

- Design and development of a set 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.

Okt. 2007

INF5220

Page #22

## 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.

## Online/Internet Field Work A definition?

---

- OWF/IWF is research into the social, cultural, political, economic, ethical, technical and aesthetic aspects of the Internet that involves observation of ongoing online events or accumulating qualitative or quantitative data from the online environments (e.g. email, web pages, discussion groups, virtual communities and/or archives) on the Internet for examination and analysis.

## Online/Internet Field Work

---

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

## 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?

## Ethical Issues, Sources:

---

- Cheltenham and Gloucester College of Higher Education: *Research Ethics: A Handbook of Principles and Procedures*.
- Association of Internet Researchers (AOIR), preliminary report on *Ethical and Legal Aspects of Research on the Internet*  
<http://aoir.org/reports/ethics.pdf>

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

---

1. Difficulty in obtaining informed consent from online subjects.
2. Difficulty of ascertaining subjects' identity because of use of pseudonyms, identity-games, etc.
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.

## Three major ethical problems

---

- 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.).

## Informed consent

---

[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, pseudonymized or anonymized data that is published) must be available for scrutiny.
- Solution(?): Retain the raw data, but pseudonymize records by using numbers instead of real IDs. Make access to RAW data very restricted (locked down - analogous to storage of sensitive data accumulated in epidemiology)

## Public or private

---

- A number of the ethical issues of covert online research disappear if online utterances are regarded as public (i.e. like books or newspaper articles) instead of private communications. What precedent (legal or otherwise) are there?
  - Against public:
    - ◆ "Glattcella" (web) as seen by Datatilsynet
  - Pro public:
    - ◆ "Synnevåg-saken" (Usenet) as seen by the police
    - ◆ "Glattcella" (web) as seen by Nettnemnda
  - Since resolved in EU court: Web pages are public.

## Institutional setting

---

- In clinical medical research, the institutional setting (i.e. the research clinic) usually have well developed procedures and mechanisms for handling, anonymizing and protecting patient data.
  - This is taken as given both by the researchers and also by the research subjects (i.e. the patients).
- In online research, no similar setting exists and has to be constructed by the researcher as part of his/her research framework.

## AOIR suggestion:

---

- Researchers need not obtain informed consent, etc., from subjects if:
  - [Prime directive:] *no intervention* with the persons whose activities are observed
  - 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  
[note: raw data should always be available for scrutiny]

## Why is online research special? Example: Handling ethics

---

Espen Munch: *En antropologisk analyse av elektronisk nettkommunikasjon*, hovedoppgave i sosialantropologi ved 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.”

Okt. 2007

INF5220

Page #37

## Pseudonymizing a direct quote

---

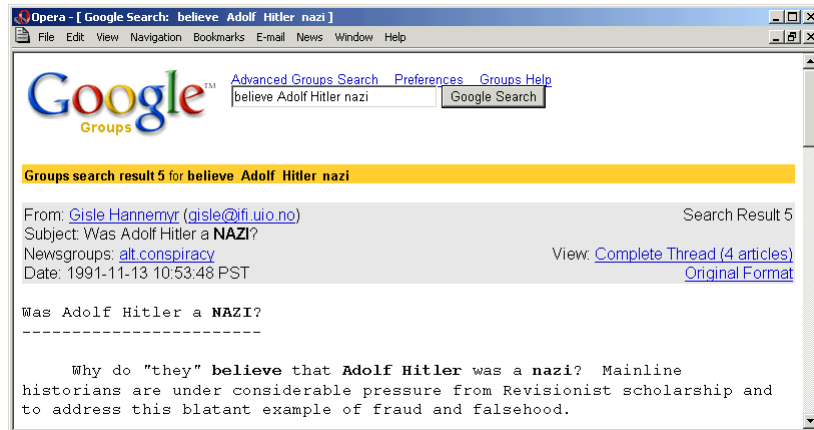
```
From: [John Doe]
Subject: Was Adolf Hitler a NAZI
Newsgroups: [some.newsgroup]
Date: [withheld]
Was Adolf Hitler a NAZI
-----
Why do 'they' believe that Adolf Hitler was a
nazi? Mainline historians are under considerable
pressure from Revisionist scholarship and to
address this blatant example of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

Okt. 2007

INF5220

Page #38

... but not very successfully




Note: Google Groups no longer reveals email address.

Okt. 2007

INF5220

Page #39

From Zack's paper

SENTENCES WITH BUSH	
<i>DRINK</i> : ACTION DONE BY BUSH	
PARTICIPANT	SENTENCE
	• <u>Bush can not drink socially because he would fall off the wagon.</u>

Anonymized participant

Okt. 2007

INF5220

Page #40

## But from Google Groups

---

```
From: "Alter" <no-e-m...@please.com>
Subject: Re: BUSH CLASSY TOWARD HILLARY; "REAL" GORE IS A CLYMER!
Date: 2000/11/06
Message-ID: <zwHN5.47178$6V6.140993@typhoon.kc.rr.com>
X-Complaints-To: abuse@rr.com
X-Trace: typhoon.kc.rr.com 973554847 24.166.255.213
        (Mon, 06 Nov 2000 17:54:07 CST)
X-MSMail-Priority: Normal
NNTP-Posting-Date: Mon, 06 Nov 2000 17:54:07 CST
Newsgroups: alt.politics.usa,alt.politics.elections,alt.politics
```

[...]

He didn't; but it's not smearing anyway. Bush can not drink socially because he would fall off the wagon.

## Final words

---

- Remember
  - Text is never just text, it is also **context**.
  - In particular, on line forums, utterances appear in a continuous stream of messages and care must be taken not to misrepresent their meaning.