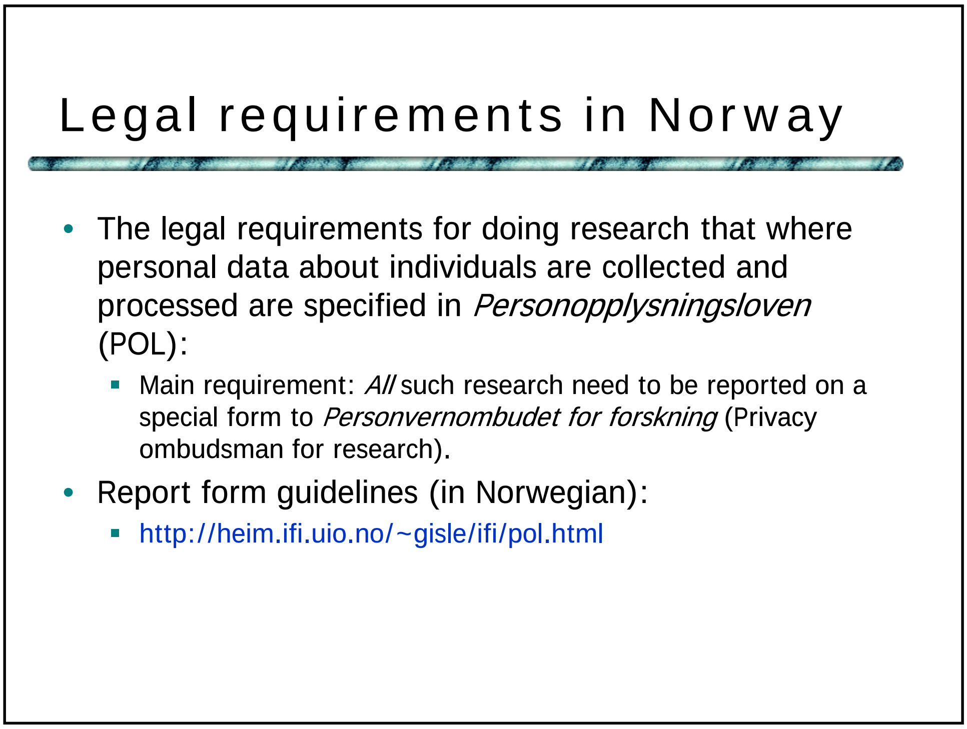
# Privacy and Internet Research (and something about citation)

Gisle Hannemyr
INF5220, autumn 2012

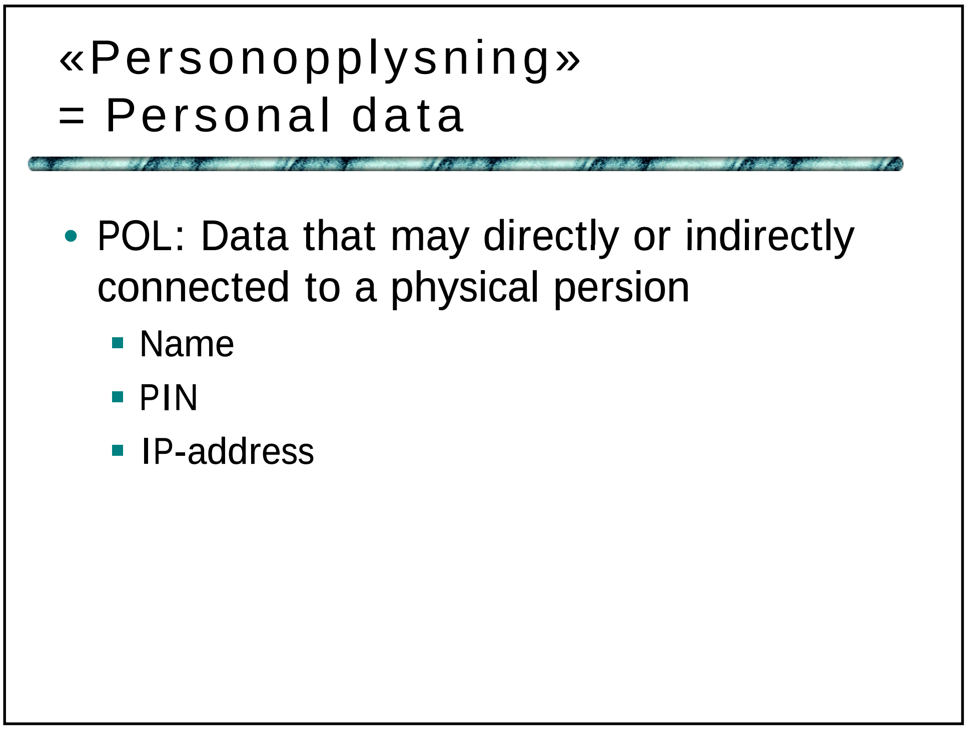
## Overview of lecture

- Legal requirements in Norway
- Internet field work
- Ethical issues
- Citing online resources



# 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



# «Personopplysning» = Personal data

- POL: Data that may directly or indirectly connected to a physical persion
  - Name
  - PIN
  - IP-address

## POL: Report form compulsory if:

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

*- or -*

- A manual register 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 insitution where the respondent is registered;
   - initiative from the respondent.
2. The responent 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.

## 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 quantitive)

## 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.

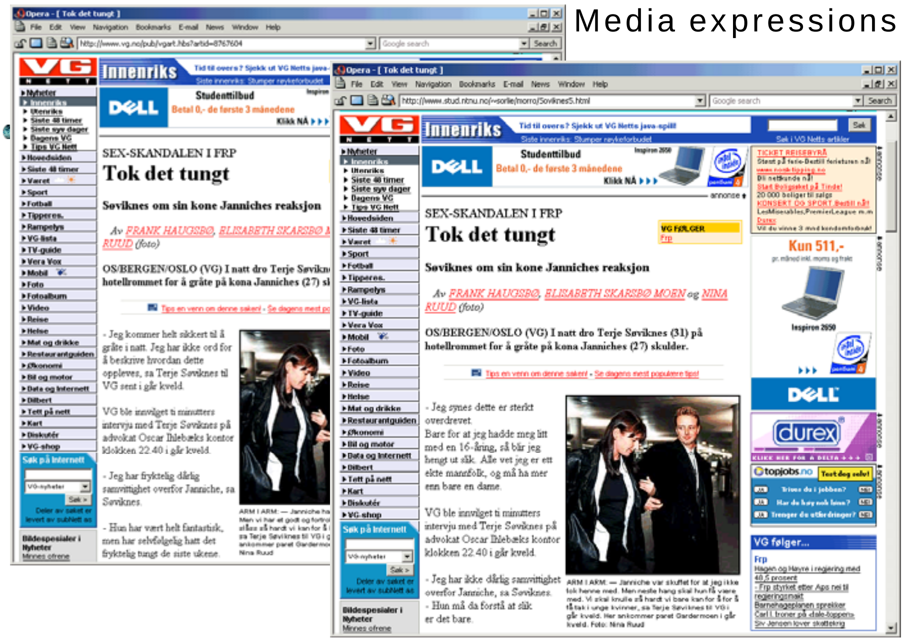
## Example: Ethnographic chat analysis

- 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 show 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.

# Media expressions



SEX-SKANDALEN I FRP

**Tok det tungt**

**Søviknes om sin kone Janniches reaksjon**

*Av FRANK HAUGSBØ, ELISABETH SKARSBØ M… RUUD (foto)*

**OS/BERGEN/OSLO (VG) I natt dro Terje Søvikn… hotellrommet for å gråte på kona Janniches (27) sk…**

Tips en venn om denne saken! - Se dagens mest po…

- Jeg kommer helt sikkert til å gråte i natt. Jeg har ikke ord for å beskrive hvordan dette oppleves, sa Terje Søviknes til VG sent i går kveld.

VG ble innvilget ti minutters intervju med Terje Søviknes på advokat Oscar Ihlebæks kontor klokken 22.40 i går kveld.

- Jeg har fryktelig dårlig samvittighet overfor Janniche, sa Søviknes.

- Hun har vært helt fantastisk, men har selvfølgelig hatt det fryktelig tungt de siste ukene.

ARM I ARM: — Janniche ha… Men vi har et godt og fortro… står så hardt vi kan for å … sa Terje Søviknes til VG i g… ankommer paret Gardermo… Nina Ruud

Opera - [ Tok det tungt ]
File Edit View Navigation Bookmarks E-mail News Window Help
http://www.stud.ntnu.no/~sorlie/morro/Soviknes5.html
Google search
Search
VG NETT
Innenriks
Tid til overs? Sjekk ut VG Netts java-spill!
Siste innenriks: Stumper røykeforbudet
Søk i VG Netts artikler
Søk
DELL
Studenttilbud
Inspiron 2650
Betal 0,- de første 3 månedene
Klikk NÅ ▸▸▸
annonse
VG FØLGER
Frp
TICKET REISEBYRÅ
Kun 511,-
Inspiron 2650
DELL
durex
topjobs.no
VG følger...
Frp

SEX-SKANDALEN I FRP

**Tok det tungt**

**Søviknes om sin kone Janniches reaksjon**

*Av FRANK HAUGSBØ, ELISABETH SKARSBØ MOEN og NINA RUUD (foto)*

**OS/BERGEN/OSLO (VG) I natt dro Terje Søviknes (31) på hotellrommet for å gråte på kona Janniches (27) skulder.**

Tips en venn om denne saken! - Se dagens mest populære tips!

- Jeg synes dette er sterkt overdrevet.
Bare for at jeg hadde meg litt med en 16-åring, så bår jeg hengt ut slik. Alle vet jeg er ett ekte mannfolk, og må ha mer enn bare en dame.

VG ble innvilget ti minutters intervju med Terje Søviknes på advokat Oscar Ihlebæks kontor klokken 22.40 i går kveld.

- Jeg har ikke dårlig samvittighet overfor Janniche, sa Søviknes.
- Hun må da forstå at slik er det bare.

ARM I ARM: — Janniche var skuffet for at jeg ikke tok henne med. Men neste hang skal hun få være med. Vi skal knulle så hardt vi bare kan for å for å tå tak i unge kvinner, sa Terje Søviknes til VG i går kveld. Her ankommer paret Gardermoen i går kveld. Foto: Nina Ruud

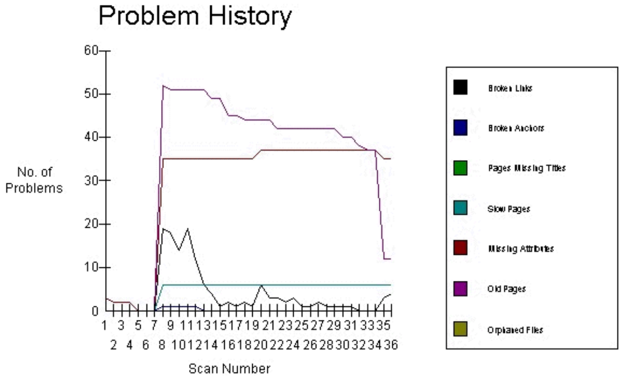
# Semi-automatic semantic analysis

- 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.

# Robot analysis
## (master thesis)

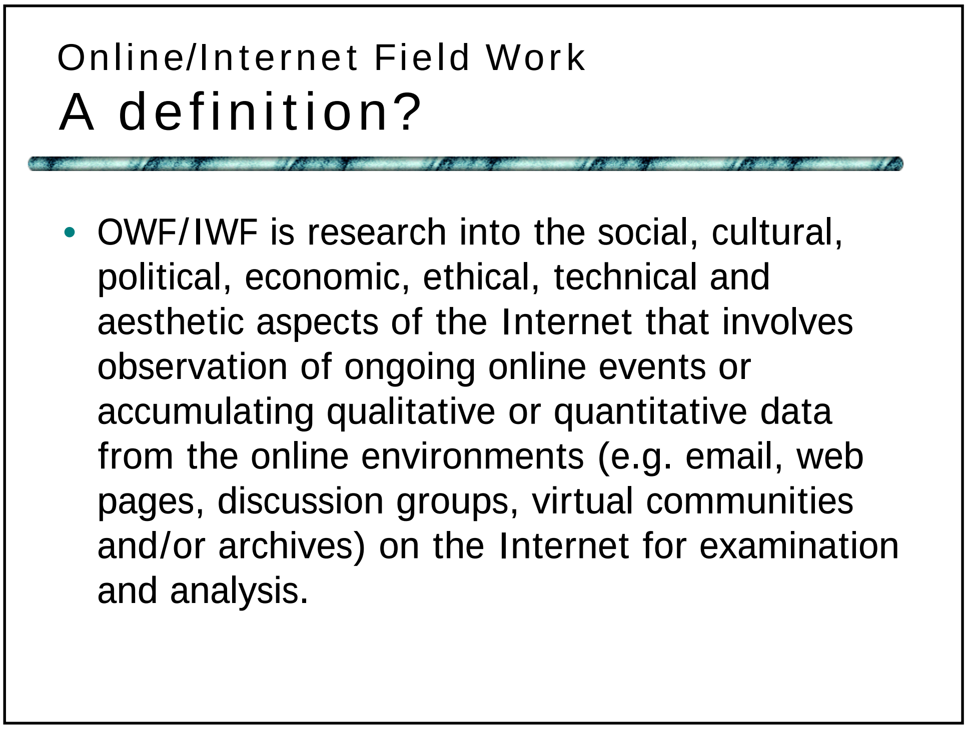
- Design and development of a set of robots and tools for analysis to measure certain aspects of the World Wide Web:
- Will accumulate data along the following axes:
  - Page size and page complexity and content (media, links, etc.)
  - Size, growth, rate of change
  - Problems: (broken links, etc.)
  - 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.

# Example:
# Robot analysis



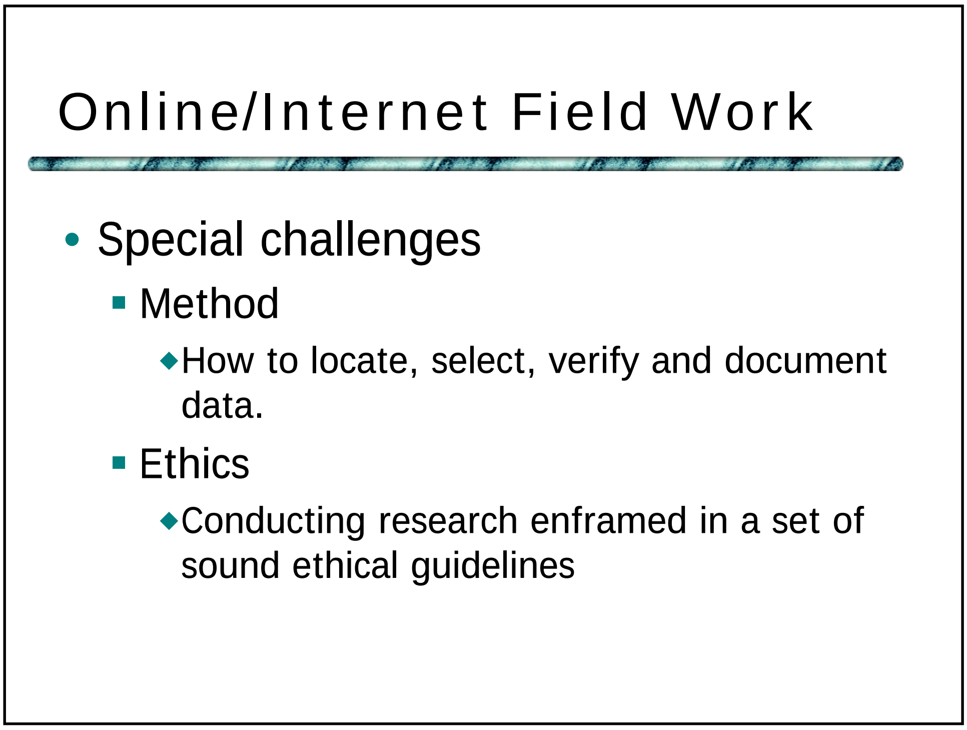
# Example:
# Robot analyzis

- 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:

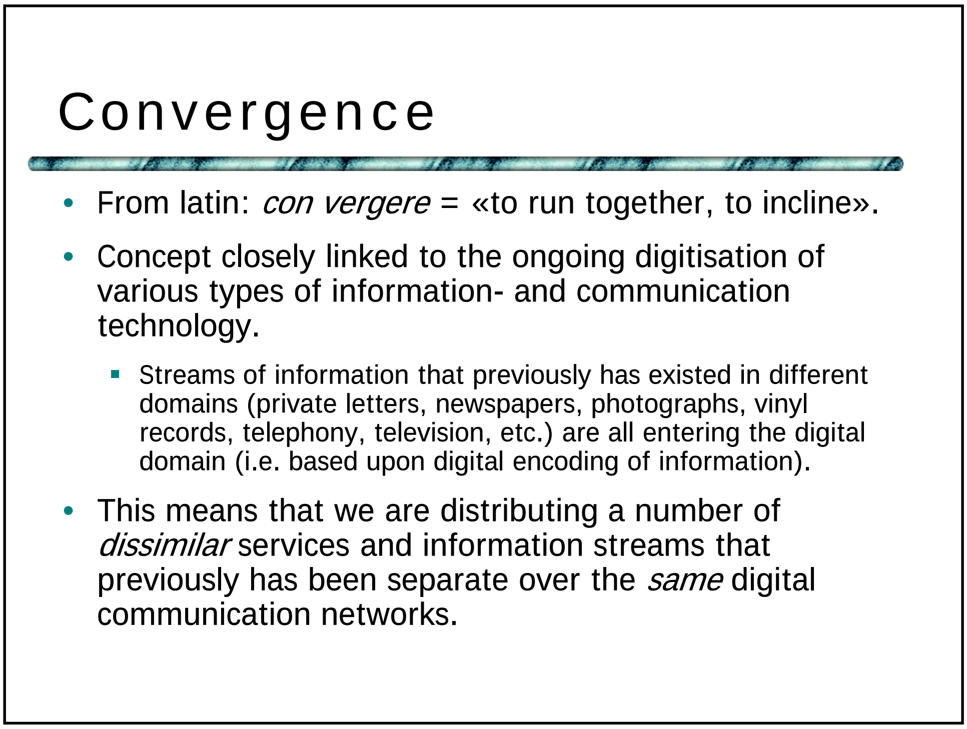
- Cheltenham and Gloucester College of Higher Education: *Research Ethics: A Handbook of Principles and Procedures.*
- Association of Internet Researchers (AOIR), preliminary report on *Ethical and Legal Aspects of Research on the Internet* http://aoir.org/reports/ethics.pdf

## 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.

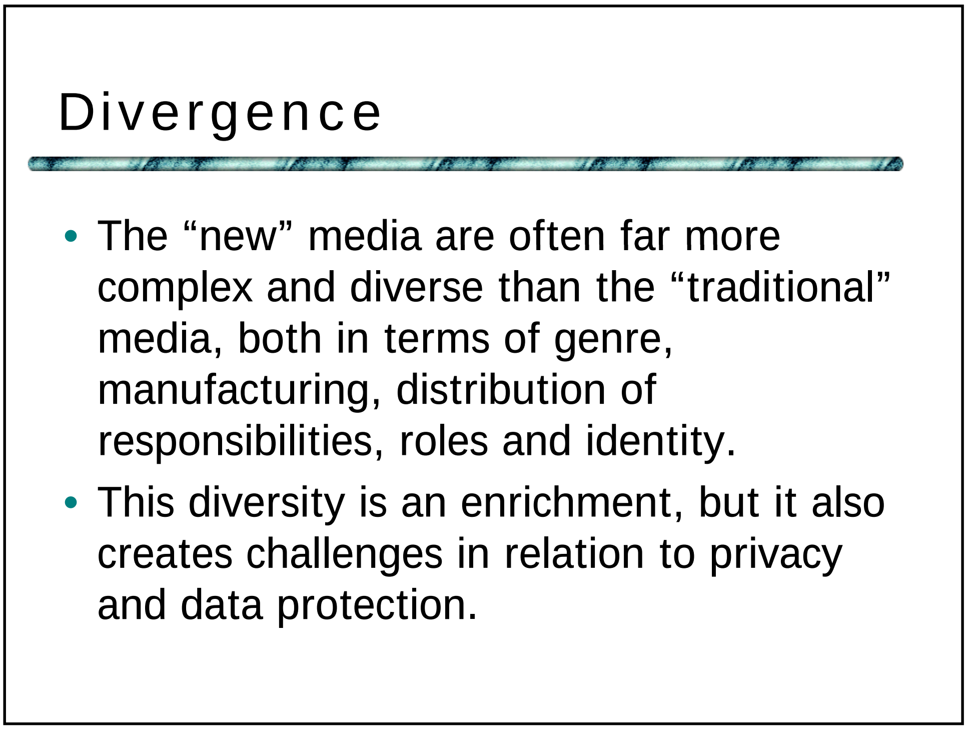
## Four major problems

- Social Media Public or Private Sphere?
- Covert research/Informed consent
- Protecting anonymity
- Raw data



## Convergence

- 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.

## 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.

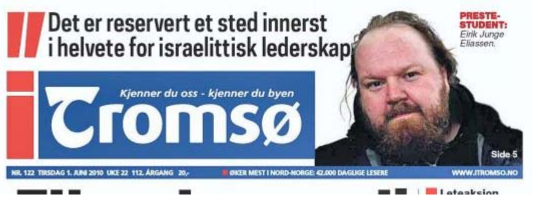
## Massemedier defined by McQuail (2000, p. 4)

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)

- Personvernombudet for research insists data data collected from interpersonal media (social media) is private.

# Facebook i bladet ”Tromsø”



Det er reservert et sted innerst i helvete for israelittisk lederskap

PRESTE-STUDENT: Eirik Junge Eliassen.

iTromsø

Kjenner du oss - kjenner du byen

Side 5

NR. 122 TIRSDAG 1. JUNI 2010 UKE 22 112. ÅRGANG 20,- ØKER MEST I NORD-NORGE: 42.000 DAGLIGE LESERE WWW.ITROMSO.NO

Leteaksjon

I kjennelsen heter det:
«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.»
iTromsø har ikke brutt god presseskikk.

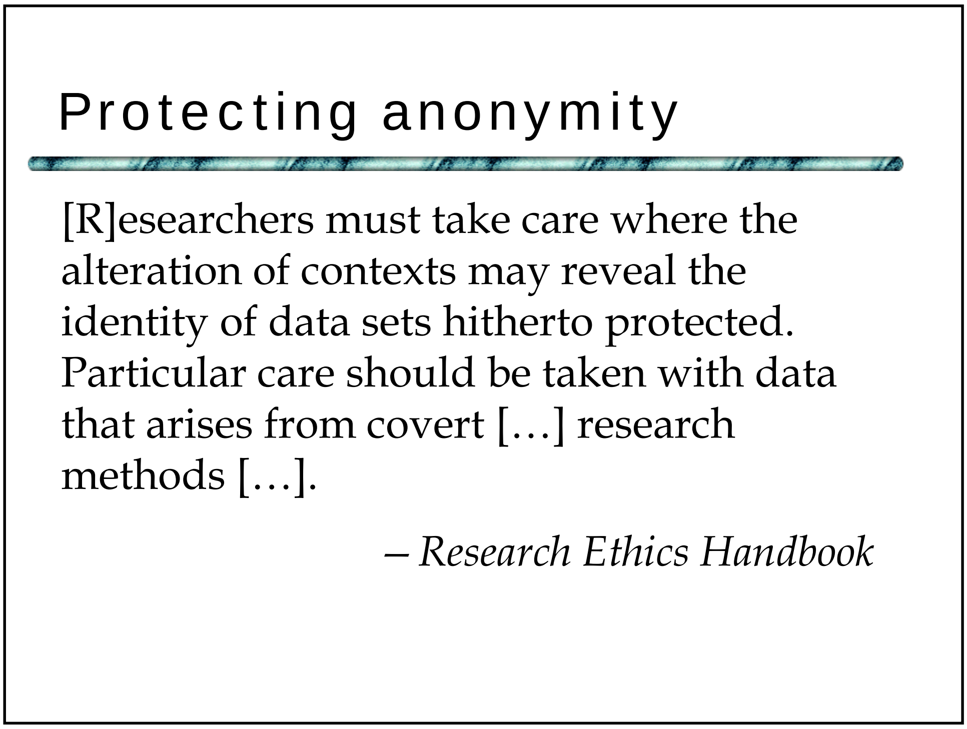
## Covert research methods

- 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

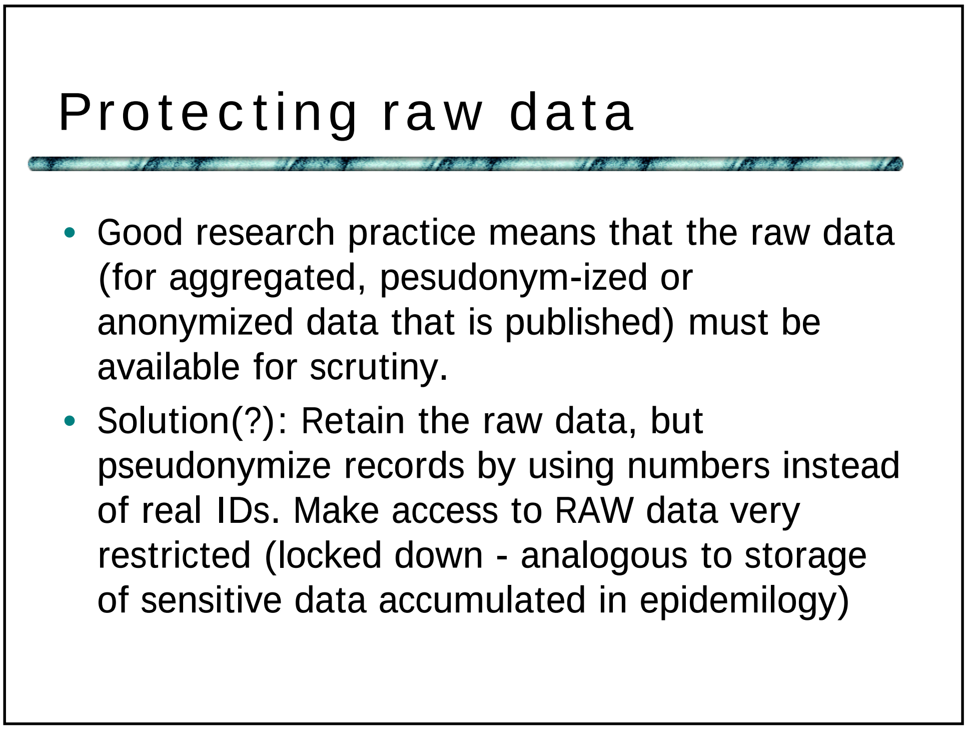
[P]rivacy is considered widely as a crucial norm in ethical research […] Data arising from research should ordinarily be considered confidential and may not be shared with others without the consent of the researched.

– *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, pesudonym-ized 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 epidemilogy)

## Institutional setting

- In clinical medical resarch, 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 resarchers and also by the research subjects (i.e. the patients).
- In online research, no similar setting exists and has to be constructed by the resarcher 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 avialable for scrutiny]

## Handling ethics: MIT “Gaydar” project

Using data from the social network Facebook, they [two MIT students] made a striking discovery: just by looking at a person’s online friends, they could predict whether the person was gay. They did this with a software program that looked at the gender and sexuality of a person’s friends and, using statistical analysis, made a prediction. The two students had no way of checking all of their predictions, but based on their own knowledge outside the Facebook world, their computer program appeared quite accurate for men, they said. People may be effectively “outing” themselves just by the virtual company they keep.

“When they first did it, it was absolutely striking – we said, ‘Oh my God – you can actually put some computation behind that,’ ” said Hal Abelson, a computer science professor at MIT who co-taught the course. “That pulls the rug out from a whole policy and technology perspective that the point is to give you control over your information – because you don’t have control over your information.”

(Source: http://scientificaesthetic.com/2009/09/24/mit-gaydar-golden-recipe/)

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

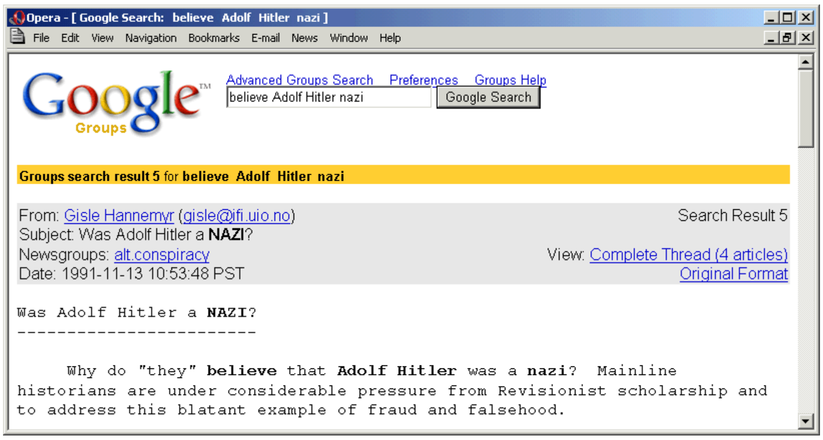
Espen Munch: *En antropologisk analyse av elektronisk nettkommunikasjon,* hovedoppgave i sosialantropologi ved UIO, 1997:

“[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.”

# 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


Opera - [ Google Search:  believe  Adolf  Hitler  nazi ]
File  Edit  View  Navigation  Bookmarks  E-mail  News  Window  Help
Google Groups
Advanced Groups Search  Preferences  Groups Help
believe Adolf Hitler nazi
Google Search

Groups search result 5 for believe Adolf Hitler nazi

From: Gisle Hannemyr (gisle@ifi.uio.no)
Subject: Was Adolf Hitler a **NAZI**?
Newsgroups: alt.conspiracy
Date: 1991-11-13 10:53:48 PST

Search Result 5
View: Complete Thread (4 articles)
Original Format

```
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.
```

Note: Google Groups no longer reveals email addess.

## 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.

## Harvard style referencing

- The preferred style of citation in information systems is a type of parenthetical referencing called “Harvard  style referencing”. It is sometimes referred to as “author-date-referencing”.
- It is believed to have originated at the Harvard University, but the best and most authoritative source to this style is the *Publication Manual of the American Psychological Association*.
- Citations are placed inside parenthesises the main body of text with the surname of author(s) and year of publication, rather than in footnotes or endnotes. The full bibliography is at the end of book or paper, sorted alphabetically on reference keyword (usually the last name of principal author).

# Harvard style (sources)

- Publication Manual of the American Psychological Association, Fifth Edition http://www.apastyle.org/pubmanual.html
- APA Style Guide to Electronic References: http://books.apa.org/books.cfm?id=4210509

# Summary of Harvard style

- Basic citation is authors last name and year in parenthesises: (Smith 2005).
- For two authors, use (Smith & Jones 2005), for more authors, use et al: (Smith et al. 2005).
- If citing  works by same author in the same year, use letters to distinguish: (Smith 2005a) and (Smith 2005b).
- To refer to a specific page in the cited work, let the page follow the year: (Smith 2006, p. 28).
- If the date of publication is unavailable, use : (Smith n.d.) (meaning: no date)

## Summary of Harvard style

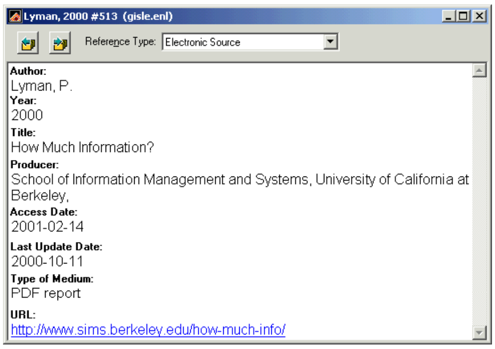
- Newspaper articles *may* be referenced giving the name and date of publication (Aftenposten Dec. 17 2005).
  - AFIK, EndNote does not support this fully automatically – must edit citation manually.
- Non-staff newspaper articles ("kronikker") should be referenced by author: (Smith 2005).
- A book published long after the original publication may be referenced as follows: (Marx [1867] 1994).
- In cases where the author is unknown:
  - If the article is written for an organization or periodical then use its name, as in (Department of Transport 2001)
  - Otherwise use the article title, italicized, as in (*Privacy in peril* 1990).
- Electronic sources are referenced just as printed sources. I'll get back to how to represent an electronic source in the bibliography.

## Summary of Harvard style

- For a quotation that is placed inline with text and marked by quotation marks, the citation follows the end-quotation mark ("), and is placed before the period:
  - "like this" (Smith 2005).
- For a quotation broken out and indented, the citation is placed after the period,
  - "like the following." (Smith 2005)
- When the author of the reference is named as part of the text itself, you may put the year only in parentheses:
  - Smith (2005) claims ...

## Online resources Maintaining bibliographies

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



Reference Type: Electronic Source

**Author:**
Lyman, P.
**Year:**
2000
**Title:**
How Much Information?
**Producer:**
School of Information Management and Systems, University of California at Berkeley,
**Access Date:**
2001-02-14
**Last Update Date:**
2000-10-11
**Type of Medium:**
PDF report
**URL:**
http://www.sims.berkeley.edu/how-much-info/

## A “better” Harvard style for EndNote (HarvardGH.ens)

Mine stiler: *http://hannemyr.com/enjoy/endnote.html*

Ifi's standard setup means that the the C-disk is only modifiable by the Administrator. To avoid being Administrator to update an EndNote library and use a custom EndNote style, keep libraries and styles below the directory: **`M:\pc\Endnote\`**, and edit your preferences: *Edit → Preferences → Folder Location → Style Folder*

## EndNote need good data

- One author per line.
- Corporate authors must end with a comma (,).
- Correct capitalisation of titles.
- Access date (and last update date if known) must be correctly entered (YYYY-MM-DD) into the appropriate fields.

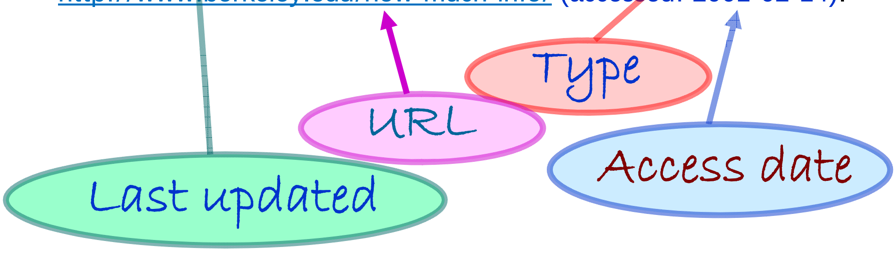
## Harvard style bibliography

Citing personal and electronic resources

Smith, W. (2002) *Citing electronic sources in scientific papers,* (private email message 2002-12-06).

Smith, W. (2003) *NCC incident,* (email interview 2003-05-13).

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


Type
URL
Access date
Last updated