



# Ethical Issues in Web Archive Creation and Usage – Towards a Research Agenda

Andreas Rauber, Bernhard Wachter  
Department of Software Technology  
and Interactive Systems  
Vienna University of Technology  
<http://www.ifs.tuwien.ac.at/~andi>

Max Kaiser  
Austrian National Library  
max.kaiser@onb.ac.at

- Web archiving is an essential activity to ensure valuable content is being preserved
- Web Archives contain a wealth of extremely valuable information

But:

- Currently most archives are closed to public
- Mostly due to legal reasons
- Need a legal solution

Is this all?

- What should such a legal solution look like?
- Is it only a legal problem?
  
- There are things that are legal, but ethically dubious
- There are things that are illegal, but ethically acceptable
  
- Privacy is an essential good
- Most countries are increasingly privacy-aware
- Are there ethical concerns, and if so
  - Are we aware of them?
  - Can we do something to address them?
- Is there a danger of a moratorium on Web Archiving?

- Hypothesis: there are a number of potentially ethically sensitive issues related to Web Archiving, and particularly to access provision
  
- In order to be able to unlock the value of Web Archives we need to
  - Understand them
  - Try to address them
  - And, provided we have appropriate technical solutions, have them reflected in the legal regulations
  
- Try to start this process via some (incomplete) considerations

- 
- Introduction: do we have ethical issues?
  - Assumptions underlying & motivating Web Archiving
  - Research questions to assist in solving a privacy dilemma
  - A quick glimpse at a case study:  
automatically identifying private content
  - Conclusions and next steps
-

## Assumptions and a number of questions:

- The Web is a new publication medium?
- The ephemeral nature of Web pages is a “design fault”?
- A Web Archive is merely a collection of publicly available information

# Assumptions underlying Web Archiving

- The Web is a new publication medium?
  - Are people “publishing”  
(conscious decision, effort invested,...)
  - If so, are they aware of it?
  - Are kids allowed to publish?
  - Which parts of the Web are publishing,  
which are communication?  
(ako chatting-in-the-bus?)
  - Do we have a choice of NOT putting some things on the  
Web?

# Assumptions underlying Web Archiving

- The ephemeral nature of Web pages is a “design fault”?
  - Post-it notes are based on a “faulty” glue  
-> should we put real glue onto them?
  - If the Web is a publication medium: may there be some who use it as such BECAUSE it is ephemeral?  
(art, temporary announcements, CV, ...)
  - Does being ephemeral make it more a communication medium in the perception of some people?
  - Does society need an ephemeral way of communicating with larger communities in an ephemeral manner?  
(speaker’s corner, graffiti, ...)

# Assumptions underlying Web Archiving

- A Web Archive is merely a collection of publicly available information
  - True, but what about Holism?  
(The whole is more than the sum of it's parts)
  - Does the ease of use, or the new possibilities of use, change the nature of an information collection?  
(full-text search, semantic analysis, IR as opposed to conventional archive catalogs)
  - Specialized person profile search engines, used by HR departments  
(special profile generation services to counter-act this)
  - Technical possibilities will increase in the future  
(video analysis, semantic analysis, reasoning, ...)

- 
- Introduction: do we have ethical issues?
  - Assumptions underlying & motivating Web Archiving
  - Research questions to assist in solving a privacy dilemma
  - A quick glimpse at a case study:  
automatically identifying private content
  - Conclusions and next steps
-

# Research questions

- What are the *ethical constraints*, and how they can be more precisely *defined or formalized*,
- Which *approaches* users of Web archives with *potentially dubious intentions* might employ to obtain information that should not be provided by privacy-respecting archives,
- In how far *technological solutions* such as query analysis, machine learning and data mining can help in identifying potentially harmful queries, potentially incriminating content on Web pages, information worth of protection, or combinations thereof,
- How *legal regulations* might be formulated in order to allow (partial) access to Web archive content in a safe, ethically correct, and useful manner

# Research questions

- Formalizing ethical constraints and risks
  - Risk analysis and threat scenarios
  - Types of information
  - Types of creators
  - Types of Web usage
  - Aggregation of information
  - Risk of NOT providing certain information

# Research questions

- Understanding potentially dubious usage
  - Who may be misusing a Web Archive?
  - For what purpose may it be misused?
  - How would the misuse look like?
  - What is the potential damage?
  - Can we detect it before misuse happens?
  - Can we put up barriers to misuse – and what would be the impact of these on normal usage
  - Is the analysis of usage ethically questionable as well?

# Research questions

- Technological approaches to solutions
  - Content mining to identify sensible information?  
(sites, pages, paragraphs, text tokens, ...)
  - Query analysis to identify patterns?
  - Site analysis to identify creators & purpose of creation?  
(children, private comment / communication, ephemeral information)
  - Technical means of controlling access and means of access?

# Research questions

- Legal solutions
  - Which legal solutions are also technically feasible?
  - How can policies be formulated, enacted and controlled?
  - How good a solution is “good enough” to be ethically acceptable, and thus legally safe to specify and act upon?

- 
- Introduction: do we have ethical issues?
  - Assumptions underlying & motivating Web Archiving
  - Research questions to assist in solving a privacy dilemma
  - A quick glimpse at a case study:  
automatically identifying private content
  - Conclusions and next steps
-

# Case Study

- Identifying potentially private information segments
- Proof of concept, meta-information
- NOT as basis for limiting/blocking access or excluding from archive, etc.
- Approach:
  - Take text documents
    - Pages
    - Paragraphs
  - Identify which ones are potentially private
  - Train a classifier (SVM, Bayesian Networks, ...)
- Similar to Genre Classification

## Two text collections:

- Santini 7-genre corpus:
  - 1400 documents (200 per genre, balanced corpus)
  - *Blog, Personal Homepage, E-Shop, FAQ, Listings, Newspaper Frontpage and Search Page*
- Product review pages
  - 50 pages amazon top-100-books 2007
  - 20 pages amazon video reviews
  - Analyzed on paragraph level:
    - description / review
    - description + professional review / “private” review
  - Using model trained on Santini 7-genre-corpus

## Features:

- Text statistics (averages)
  - # words, +syllables, word-length
- Tokens (rel. frequ.)
  - word-, symbol-, space-, number-, punctuation-, control-tokens
- Readability Indices
  - Flesh Reading Ease, Flesh-Kincaid Grade Level
- Look-ups (rel. frequ.)
  - places, time/date, person names, currencies, addresses, company names, abbreviations, telephone/fax numbers
- Part-of-Speech (rel. frequ.)
  - nouns, verbs, adjectives, adverbs, modal verbs, conjugations, pronouns, personal pronouns, articles, prepositions, exclamations, list elements
- Presentation (rel. frequ.)
  - links, headings, paragraphs, text formatting, lists, tables, graphics, frames, forms, multimedia elements

# Case Study

## Results:

- Classification Santinin 7-genre corpus using SVM

Genre	Precision	Recall	F2
Blog	93,17	87,22	90,01
Homepage	89,59	48,33	62,79
Private total	91,38	67,78	76,40

- Classification product pages, Naïve Bayes, 7-g model

Corpus	Precision	Recall	F2
top-100-books	77,91	84,07	80,87
top-100-books-b	78,27	85,75	81,84
video games	87,66	96,02	91,65

# Case Study

## Results:

Emily, 22, an assistant in nearby Stone clothes shop, witnessed the aftermath. She said: "There were two guys lying on the floor who had been stabbed. They looked like they were in a lot of pain."

A family friend, who asked not to be named, said the murdered teenager was a "good boy who would never hurt anyone".

She said: "I have known Naz since he was a toddler. When boys get murdered sometimes you hear bad things about them but you would never hear anything bad about Naz. He was loved by the community, he was a peacemaker. He was a good family boy."

Speaking outside the terrace Victorian family home in Highbury, his older brother said: "He was a very good boy, loved by everyone. "He was doing his A-levels and never in trouble. We are just shocked and saddened by this. We can't really speak about it now."

An 18-year-old man was arrested in the area following the lunchtime attack. He remains in custody in a north London police station.

**Share this article:**  
[What is this?](#)

- [Digg it](#)
- [Del.icio.us](#)
- [Reddit](#)
- [Newsvine](#)
- [Nowpublic](#)

**Comment Add your comment Comments (20)**

20 people have commented on this story so far. Tell us what you think below.

Another murder in Nu Lab's Britain. And once again the scum have been allowed to kill a decent person! Will this useless government ever wake up to the problems facing this country?

- David Simpson, Heckmondwike ← probability: 0.95

No doubt these gangs must be pandered to and their human rights respected. ← probability: 0.87

- Phil Davy, Eltham, London ← probability: 0.82

Yet another murder. Deepest condolences to this young man's family. One of these days we will have a party in power who say "to hell with these rights, those rights etc let's get back to old fashioned rules and values" and I will be the first to have helped vote them in! ← probability: 0.97

# Conclusions

- Web Archives contain valuable information that would be very useful to make freely available to the public
- We do believe there are ethical issues with (providing access to) Web Archives
- We need a legal solution to these problems
- For a legal solution we need an understanding of
  - The problems and challenges
  - Risks and their consequences
  - Technical means to automatically counter these and to enforce legal regulations
- Otherwise we may face the risk of having a moratorium / legal regulation stopping Web Archiving all together....  
....or at least feel very bad continuing doing it

# Conclusions

- **What we want are**
  - privacy-aware
  - ethically responsible
  - comprehensive
  - powerful and
  - freely available

Web Archives to serve the needs of society now and in the future

- This is not to stop / hinder Web Archiving, but to help it to benefit society and to become / remain useful
- Comments, feedback, directions & discussion are more than welcome and needed!