Workshop on Web Archiving

MODULE 1 A:
WEB ARCHIVING: Theory - and a Bit of Practice

Niels Brügger
Asger Harlung
Program 16.01.2018, KU

11.00-12.30  Workshop part 1: Web Archiving Theory

12.30-13.15  Lunch

13.15-13.45  Workshop part 2: Existing collections

13.45-15.00  Workshop part 3: Doing your own archiving
             (Short introduction, after that individual breaks as needed)

15.00-15.15  Wrap-up; feedback
Module 1: Web Archiving

- Introducing ourselves and NetLab
- Why archive the web
- Research examples
- Three kinds of digital content
- WWW as technology
- What is web archiving?
- Methods of web archiving
- Challenges for the web crawler
- Crawling — advantages/disadvantages
- Characteristics of the archived web
Niels Brügger – Professor (MSO, with special responsibilities) in Internet Studies and Digital Humanities, Head of NetLab, and of the Centre for Internet Studies, specialising in internet research since 1997.

Asger Harlung – MA in ICT and learning, has previously worked with research in digital rhetoric, and supporting creativity development in learning processes.
A research infrastructure for internet research.

Part of the Danish research infrastructure Digital Humanities Lab (DIGHUMLAB).

Established in 2012.

Research driven development of research infrastructures.
Digital Humanities Lab

Language Tools (KU)
Media Tools (AU)
Interaction & Design (AAU & SDU)

NetLab
Audio and visual materials

Online
Archived
IT architect
Netarkivet (the Danish national web archive)

NetLab Forum

Collecting data for specific projects
NetLab Services

NetLab’s services are free for members of the DIGHUMLAB communities (AU, AAU, KU, SDU, KB)

Our services include:

• Website (free and open resource)
• Tools and tutorials
• Workshops (on demand and on site)
• Online courses
• Advice and support for research
• Research assistance (upon application, call coming in 2018)
Three Statements

- 2000: 75% of the world’s data was stored in analog form (paper, film, photographic prints, vinyl, magnetic cassette tapes, etc.),
- 2007: 7% analog, 93% digital
- 2012: Only 2% of all stored data was stored in analog form.

Mayer-Schönberger & K. Cukier (2013):
[...] the human demand for more comprehensive digital memory will continue to rise. The result is a world that is set to remember, and that has little if any incentive to forget.

Three Statements

• [...] it takes about 50 days for 50% of the web to change or to be replaced by new pages (Cho and Garcia-Molina, 1999, p. 7).

• The survival survey revealed that more than 90% of the web pages had disappeared in the last 12 years. The life span study found that the average life span of a web page is 1,132.1 days (Agata et al., 2014, p. 464)

• 50% of resources [are] unrecognisable or gone after 1 year, 60% after 2 years, 65% after 3 years (Jackson, 2015, p. 20)
Why Archive the Web?

- To preserve the cultural heritage
- To preserve a stable research object
- To be able to document and illustrate a study
- Modern source references
- Documentation in general; legal claims
The Research Process

Corpus creation
- Search
- Identify

Analysis
- Duplicates
- Evaluate
- Select
- Select/remove/combine

Dissemination

Storage

Close — middle — distant reading
dr.dk — FV11-15 — entire .dk
The Research Process

- Corpus creation
- Analysis
- Dissemination
- Storage

Close — middle — distant reading
dr.dk — FV11-15 — entire .dk
The Research Process

Corpus creation → Analysis → Dissemination → Storage

Referencing

Close — middle — distant reading
dr.dk — FV11-15 — entire .dk
The Research Process

- Corpus creation
- Analysis
- Dissemination
- Storage

Close — middle — distant reading

dr.dk — FV11-15 — entire .dk
The Research Process

WORKSPACE

Corpus creation

DOCUMENTATION

Analysis

Dissemination

DATA MANAGEMENT

LEGAL ADVICE

Storage

DOCUMENTATION

DATA MANAGEMENT

LEGAL ADVICE
Case 1: Harvesting the Digital Music Revolution

Henrik Smith-Sivertsen
Project researcher, The Royal Library
Denmark

Research topic:
Archiving the digital Music Revolution
– The Case of Sys Bjerre
Results:
• The initial study showed that without a search function, it was almost impossible to use and evaluate Netarkivet.
• After presentations of the results, the NetLab group pushed for a search option.
• When a search option was implemented, a comparison between the initial results and key word searches showed significant differences.
Case 1: Harvesting the Digital Music Revolution

Through Netarkivet we were able to study how Sys Bjerre was marketed as a DIY Artist in the founding years of her career, her changing strategies and how the web was fundamental to her.

After full text search was implemented, we were even able to study how a viral hit was spread in 2008.

By 2017 most of these sources have been deleted online.
Case 1: Harvesting the Digital Music Revolution

Publications and presentations:


Case 2: The New Nordic Fatherhood

Stinne Krogager
Associate Professor in
Communication, Aalborg University

Research topic:
The New Nordic Fatherhood

Research context:
Article for an upcoming anthology in
the book series:

MÆRKK – Æstetik og Kommunikation
at the publisher Systime.
The article will present an analysis of the phenomenon of fathers embracing cooking over the last decade to a degree and extent that is unique to Denmark, and how this defines fatherhood,

a) As an engaged and household contributing father role,

b) As an aethetical expression of fatherhood.

The analysis will take into account:
Representation of the trend of coocking fathers in social media, and also in a contextualised in perspective of historical aesthetic representations of fatherhood and masculinity ideals in Denmark.
Case 2: The New Nordic Fatherhood

The article will present an analysis of the phenomenon of fathers embracing cooking over the last decade to a degree and extent that is unique to Denmark, and how this defines fatherhood,
a) As an engaged and household contributing father role,
b) As an aesthetic expression of fatherhood.

The analysis will take into account:
Representation of the trend of cooking fathers in social media, and also in a contextualised in perspective of historical aesthetic representations of fatherhood and masculinity ideals in Denmark.
Thomas Madfar 🍩 madfar.dk 🌐 mad@madfar.dk 🍴 medejør Café Lindholm #cafelindholm 🍽 kogebog - Madfar inspiration til dit hverdagskøkken #madfar madfar.dk
Case 3: Syrian Interest Groups

Ally McCrow-Young
PhD fellow, University of Copenhagen

Research topic:
Syrian Interest Groups’ Activities on the Web

Research context:
Research group: Images of Conflict, Conflicting Images

Funded by the Velux Foundation
My research project analyses visual social media use which counters Isis through image creation and dissemination.

My area of interest regarding web archiving is to preserve and document digital images that spread across various social media platforms.
Case 3: Automated Archiving Tools

- **Google Reverse Image scraper**
  https://tools.digitalmethods.net/beta/googleReverseImages/#
  - Successful for analysing the distribution of specific images

- **Stillio** (https://app.stillio.com)
  - Useful for regular screen captures of webpages, also trialed with Twitter users

Example of Stillio screenshot command: set to create daily archives
Case 3: Supplemental Manual Archiving

• FireShot
  Successful for archiving scrollable Twitter ‘media’ streams

• Video download helper
  To supplement Twitter capture through FireShot

• Chrome export to PDF
  Successful for Instagram archiving (based on both hashtags and individual users)

Left: Example of FireShot screenshot capture tool.

Right: Example of Chrome export to PDF from Instagram.
Probing a Nation’s Web Domain — from Small Data to Big Data
The historical development of an entire national web: .dk 2005-2015

The project is a collaboration with Netarkivet.
Brutto list of 'probes':

• Size — e.g. bytes
• Space — e.g. geolocalisation
• Structure — e.g. network of hyperlinks
• Liveliness — e.g. domain names and updating
• Content — e.g. degrees of openness, files, software types, language, website textual elements, semantics
Digitised
Formerly analog media, transferred to a digital form.

Born Digital
Has not previously existed in any other form than digital.

Reborn Digital
Born digital content which has been gathered and preserved, and to some extent has been changed in the process.
WWW — one among other internet protocols:

http — Hyper Text Transfer Protocol
URL — Uniform Resource Identifier (Locator)
html — Hyper Text Markup Language

Constructing a URL on WWW:

protocol://subdomain.domain.topdomain/path/page/
http://cc.au.dk/research/researchprograms/
Web pages = patched together in an ‘empty’ shell (stylesheet) of material from databases

The browser (Safari, Firefox...) translates html into writing, pictures etc.
Small Exercise: Source Code

Snagit

Snagit is a tool for screen capture and also be used to capture an individual webpage.

The application is available for $49.95 and allows for a fifteen days trial period.

https://www.techsmith.com/snagit.html

Works on:

![Windows Logo]

```html
<html>
  <head>
    <title>Example HTML Page</title>
  </head>
  <body>
    <!-- HTML code goes here -->
  </body>
</html>
```
What is Web Archiving?

International Internet Preservation Consortium’s definition:
”… the process of gathering up data that has been published on the World Wide Web, storing it, ensuring the data is preserved in an archive, and making the collected data available for future research.”

(http://netpreserve.org/about-us)

”Any form of deliberate and purposive preserving of web material.” (Brügger, 2011:25)

What is Web Archiving?

Macro archiving
• Cultural heritage institutions
• Preserve as much as possible
• Big and varied data
• IT expertise, advanced technology, computer power

Micro archiving
• Individual researcher/research group
• Stabilize a concrete research object, here-and-now
• No experience, no advanced technology or computer power
Methods of Web Archiving

• Web crawling (hyperlink crawling)
• Screen image
• Screen filming
• Harvesting via API
• (Delivery from producers)
Web Crawling
Web Crawling

crawler

---

domain.com

1. page
2. page
3. page

page

---

Workshop KU
16.01.2018

netlab.dk
Web Crawling

By-Harvest
Challenges for the crawler

- JavaScripts
- Content based on Flash
- Interactive pages
- Streamed content
- Websites with access limitations (password, captcha)
- Cookies, adds, plugins etc.
- Robots.txt
- Deep web (e.g. databaser, ftp-server, password-protected content, hidden content, pages not linked to, dynamic content based on requests).

http://da.wikipedia.org/wiki/CAPTCHA
Pages not being crawled

- Not crawled – robots.txt
- Not crawled – password protected
- Not crawled – script
- Not crawled – too deep

NetLab
netlab.dk
Elements not crawled

Netarkivet
Elements not crawled

Netarkivet
Elements not crawled

Internet Archive
Crawling, Advantages

• The entire page in full length
• Hyperlinks, link source as well as target
• Look and feel of live web (with limitations)
• Automatic (partly, evaluation and trouble shooting)
• Machine readable, enables search, sorting, analysis
• Access to metadata (crawl logs)
• Robust format (html)
• Big data-analysis (content analysis, network analysis, etc.)
Crawling, Disadvantages

• Some objects not archived, e.g. videos and streamed content, and applications based on Flash, JavaScript etc.

• Temporal inconsistencies

• Difficult to delimit in terms of spatial extent

• Risk of web crawler being caught in ’bot traps’ (some monitoring is necessary)
Characteristics of the Archived Web

What is archived is not a 1:1 copy of the material one attempted to archive

It is versions/reconstructions:

- Created in the process of archiving
- On the basis of a number of choices made by the archiver (harvesting strategy, settings, etc.)
- The choices made have consequences for what is archived
- The archived objects are re-assembled in the archive → ’replay’
The problems that regular web pages present to web crawlers are significantly multiplied by the functions and characteristics of social software.

We will leave the presentation for short while in order to demonstrate some solutions that may help in individual cases.
Characteristics of the Archived Web

The archived version is deficient because of:

• Technical challenges
• Web’s specific characteristics: dynamic, unpredictable
• Potential asynchronicity between updating and archiving

→ archiving takes time
→ certain elements cannot be archived

It is an added challenge that we do not know what is missing:

• Not much documentation
• No baseline to compare with
Characteristics of the Archived Web

As scholars using archived web as an object of study, it is important that we are aware of the pitfalls and sources of error inherent in the material.
Characteristics of the Archived Web

It is versions/reconstructions:

- The archived objects are re-assembled in the archive → 'replay'
IN CONTRAST TO DIGITIZED COLLECTIONS: TO A LARGE EXTENT ARCHIVED WEB IS ALREADY MARKED UP — HTML, FILE NAMES...