Results from exploring the archived web
Three Statements

2000: 75% of the world’s data was stored in non-digital form (paper, film, photographic prints, vinyl, magnetic cassette tapes, etc.),

2007: 7% non-digital, 93% digital

2012: Only 2% of all stored data was stored in non-digital form.

Three Statements

[...] 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)
A research infrastructure that can help researchers unlock and study this treasure trove of digital material — including the web.
NetLab

NetLab is a national digital research infrastructure to support research in the online and the archived web.

NetLab offers a wide range of services, including:
• Tools and tutorials
• Workshops
• Courses, including online courses
• Events in NetLab Forum
• Advice and support for research

Close collaboration with Netarkivet at the Royal Library
Necessary Scholarly Skills

Archived web differs from other types of digital collections.

Archived as bits and pieces — can be taken out of the web archive and recombined in a great variety of ways.

Knowledge and methodological skills as to how to use the archived web as a research object are needed.
Six Cases

Cases of contemporary research in topics where archived web content was/is a mandatory resource.

All researchers in the following cases have at one or more occasions been supported by NetLab in their projects.

Invited to send us slides — slides shown as received.
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
Case 1:
Harvesting the Digital Music Revolution

Main goal:
Evaluation of Netarkivet as a resource for recent music history

Method:
Lab experiment using Netarkivet as the only source.
Case 1: Harvesting the Digital Music Revolution

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: Social Media as Semiotic Technology

Søren Vigild Poulsen, Assistant Professor, University of Southern Denmark
Gunhild Kvåle, Agder University

Research topic:
Social Media as Semiotic Technology: A Multimodal Social Semiotic Perspective
Case 2: 
Social Media as Semiotic Technology

Research context:
Article with working title:
"Building meaning potential – historical study of Instagram’s user interface design"

For submission to a journal such as, Visual Communication, or Internet Histories
The historical interest in social media involves several dimensions:
• historical development of a given social medium
• social medium’s connections to previous media
• the role of social media in transformations of existing social practices and the emergence of new practices
• discourses that surround social media
Historical Perspective  
(Poulsen, in progress)

### Instagrams blogarkiv

Symbols: α = addition, β = elaboration, _ = no addition or elaboration, status quo

Categories of tools: User Interface, Profile, Take photo, Editing, Sharing

<table>
<thead>
<tr>
<th>X-axis: Versions</th>
<th>1.0</th>
<th>1.1</th>
<th>1.7</th>
<th>1.8</th>
<th>1.8.5</th>
<th>2.0.1</th>
<th>2.0.2</th>
<th>2.1</th>
<th>2.1.1</th>
<th>3.1</th>
<th>3.2</th>
<th>4.0</th>
<th>4.1</th>
<th>4.2</th>
<th>4.3</th>
<th>4.4</th>
<th>4.5</th>
<th>5.0</th>
<th>5.1</th>
<th>5.2</th>
<th>5.3</th>
<th>5.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y-axis: Tools</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User Interface</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Home/Feed</td>
<td>α</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>β</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Popular/Explore</td>
<td>α</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>β</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Take photo</td>
<td>α</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>β</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>News/Follow</td>
<td>α</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>β</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Profile</td>
<td>α</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>β</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Username</td>
<td>α</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>β</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Profile photo</td>
<td>α</td>
<td>β</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>β</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Profile Bio</td>
<td>α</td>
<td>-</td>
<td>β</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>β</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Grid/List Photo</td>
<td>α</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>β</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Push notification</td>
<td>α</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>β</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Photos you liked</td>
<td>α</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>β</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Take photo</td>
<td>α</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>β</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Take photo (incl. camera)</td>
<td>α</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>β</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Live video</td>
<td>α</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>β</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Image library</td>
<td>α</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>β</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Editing</td>
<td>α</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>β</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

11/13/2017
Digital Humanities Lab Denmark
Changes in tools:  
Semiotic representation of a tool,  
e.g. Take photo  
(Poulsen, in progress)

Version 1.7.

Version 1.8.

Version 2.1.

Version 5.1. (Android)
Case 2:
Social Media as Semiotic Technology

Learned about archived content and archiving tools at NetLab’s Workshop
Consulted NetLab for additional advice
Case 3:
The New Nordic Fatherhood

Stinne Krogager
Associate Professor in Communication,
Aalborg University

Research topic:
The New Nordic Fatherhood
Case 3: The New Nordic Fatherhood

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

MÆRKK – Æstetik og Kommunikation

at the publisher Systime.
Article: 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 coocking fathers in social media, and also in a contextualised perspective of historical aesthetic representations of fatherhood and masculinity ideals in Denmark.
Case 3: The New Nordic Fatherhood

Learned about archived content and archiving tools at NetLab’s Workshop
Consulted NetLab for additional advice

“NetLab has provided support on more occasions, when applications encountered limitations that I did not have sufficient knowhow to work around. With NetLab’s help I managed to archive the data material using a number of different applications.”

- Stinne Kroager
Case 4: Syrian Interest Groups

Ally McCrow-Young
PhD fellow, University of Copenhagen

Research topic:
Syrian Interest Groups’ Activities on the Web
Case 4: Syrian Interest Groups

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.
Automated Archiving Tools

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

- **Stillio** ([https://app.stillio.com](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
Supplementing Automated Tools with 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.
Case 4: Syrian Interest Groups

Received project-specific feedback and learned about archived content and archiving tools at NetLab’s Online Course on Web Archives and Web Archiving.

“Thanks for an interesting and practical course. It has certainly given me a lot to think about, and many useful ways of approaching my data gathering for my project!”
– Ally McCrow-Young
Case 5:
The Danish HPV Vaccine Controversy

Marie Louise Tørring, Associate Professor, Aarhus University
Stine Hauberg Nielsen, PhD Fellow, Aarhus University

Research topic:
The Danish HPV Vaccine Controversy
Case 5:
The Danish HPV Vaccine Controversy

Research context:
Public Health at a Crossroads – PUSH

Research stipend from the Danish Cancer Society:
“When vaccines go viral: A study of the role of social media in spread of HPV vaccine anxiety in Denmark”
In 2015, Denmark saw a drastic increase in reported side effects from the HPV vaccine and at the same time a drastic decrease in public commitment to the HPV vaccination programme.

Disputes concerning the HPV vaccine have become a fertile soil for a heated controversy in different areas of Danish society. This development is undoubtedly driven by a new digital media reality, which decentralises and pluralises health information and puts classical preventive health initiatives under pressure.

Source: The Danish Health Authority and The Danish Medicines Agency
From an anthropological perspective, the HPV vaccine controversy is an extremely interesting and relevant case, but it is also challenging, both in respect of theory and methodology. There are several problem areas that are relevant to study in connection with the HPV vaccine controversy, but which are also difficult to study solely by means of ethnographical methods. Examples of such problem areas are:

- **Scale**: How different aspects of the HPV vaccine controversy are represented and connected across country borders,
- **Complexity**: How different players, interests and focus areas interact in this field,
- **Dynamics and rapidity** – how the controversy develops and transforms with accelerating speed and unforeseeability.
In our attempt to account for these challenges, we are using:

- **Digital mapping** of the HPV vaccine controversy, allowing us to map, analyse and graphically visualise elements such as, linguistics, players, networks, connections, and temporal dynamics.

- **Archived web content**, allowing us to study the HPV vaccine controversy in retrospect and analyse what went on before and during the controversy breakout. With archived web content we have access to studying the HPV vaccine controversy and its development as it occurred; in its own "real time".
Case 5:
The Danish HPV Vaccine Controversy

Learned about archived content and archiving tools at NetLab’s Workshop

“NetLab’s workshops ought to be a mandatory methodology course for all Master’s programs in the Humanities department.”
– Marie Louise Tørring
Case 6:
The Danish Web Domain 2005-2015

Niels Brügger, Professor (MSO), Aarhus University
Janne Nielsen, Associate Professor, Aarhus University
Ulrich Have, IT Architect, NetLab
Ditte Laursen, The Royal Library
Per Møldrup-Dalum, IT developer, The Royal Library

Research topic:
The historical development of the Danish web domain .dk 2005-2015
Case 6:  
The Danish Web Domain 2005-2015

Research context:
Pilot project at the DeIC Cultural Heritage Cluster at the Royal Library (HPC)
Grant from the Ministry of Culture
Aarhus University & The Royal Library/Netarkivet

Based on material from the national Danish web archive Netarkivet
Case 6: The Danish Web Domain 2005-2015

The historical development of an entire national web .dk 2005-2015

- Size and numbers — e.g. bytes, number of files
- Space — e.g. geolocalisation
- Structure — e.g. network of hyperlinks
- Liveliness — e.g. domain names and updating
- Content — e.g. degrees of openness, software types, language, textual elements, semantics
Case 6:
The Danish Web Domain 2005-2015

Finally the Cultural Heritage Cluster is working!

The very first test results were there Tuesday last week — years 2006 and 2014.

Disclaimer: “This is a first attempt to extract data from the Netarkiv and test the performance of the system. The data extracted will be presented as is without many considerations of its validity nor any amount of quality assurance.”
Top 18 of harvested MIME types in 2006 (same scale as for 2014)
Case 6: The Danish Web Domain 2005-2015

Top 18 of harvested MIME types in 2006 (same scale as for 2014)

Top 18 of harvested MIME types in 2014

Digital Humanities Lab Denmark
Thank you. Visit us at netlab.dk