Social media archiving at the Royal Danish Library

– Reflections and status quo

Anders Klindt Myrvoll, Programme Manager Netarkivet, the Danish web archive
Digital Cultural Heritage

Wanted: Social Media Data – Archiving Practices and Research Use

15. September 2022 – KBR Bruxelles
AGENDA

• Background & legal deposit law
• Accession- / crawl strategy
• Status quo + stats
• A way forward
• Search/discovery & playback
The Black Diamond in Copenhagen – photo: 2018 Laura Stamer
What the legal deposit law says

- **July 2005**: a revision of the Danish legal deposit law comes into force

- **Danish content** published in electronic communication networks
  - Published by Danes
  - Published in Danish
  - Adressed to a Danish audience

Translation of Act No. 1439 of 22. December 2004

*Unauthorised version*
1697: First legal deposit law
(KB – only printed media)

1902: Revision
(KB + SB – only printed media)

1998: Revision
+ CD's and video
+ "static" internet

2005: Revision + radio/tv + "dynamic" internet + film

"Material published on electronic communications networks"
Link graphs

34 TB data
Annually harvested. Preserved in newly developed Bitmagasin-

~ 2.7 mio.
domains
harvested

~ 37,005,089,197 objects

content_type_norm
• html (23,249,386,801)
• image (6,918,535,128)
• other (5,263,047,834)
• text (1,442,841,304)
• pdf (111,443,762)
• audio (10,311,066)
• video (9,459,710)
• word (51,322)
• powerpoint (6,952)
• excel (5,318)

more content_type_norms
Crawl strategies

Broad crawl
Top level domain. All .dk-domains 3-4 times a year.

Selective harvests
All Danish news media (e.g. Daily, weekly, max. 12 times a day), political parties, organizations and associations, ministries and agencies, selected profiles from social media

Event harvests
1-3 events yearly, e.g. Corona/Covid 19, elections.

Special harvests
Closing of sites (e.g. Google+), Flickr (changing terms/potential loss of data), requests from researchers
Social Media strategy

• There are particular challenges in collecting material via publicly available profiles on social media.

• Therefore, the following services are selected in order of priority: Twitter, Facebook, YouTube, Reddit, Instagram, TikTok, Tumblr and other media that appear

• Technical restrictions (sites exclude "robots/harvesters/crawlers")

• Difficult to identify all relevant material in an automated (as well as manual) way

• Storage / data quickly becomes bulky. We can not collect everything.

• Dynamic content, embedded files, and dynamically generated URLs is not easy to crawl
Social Media strategy

The Royal Danish Library's web curators maintain a representative selection of SoMe profiles and pages in consultation with the library's other curators.

The Royal Library will collect profiles and pages related to Danish news, social affairs, politics and culture, such as politicians, opinion makers and frequent publishers on the selected social media, who have high numbers of followers, many comments on the content or subscribers.

Before the 2019 general election we made a big list of 1800+ Facebook profiles and tried to get this content via Archivelt.

We identified all 800+ politicians that ran for parliament and researched to find more of the most influential Facebook profiles, political parties etc.

It’s ongoing work to update this list and we’ll iterate this autumn before the coming general election.

We have found some ranking-pages, but we’ll look further into the subject this fall – this could be business intelligence solutions, top lists etc. combined with our own curated lists.
Social Media strategy – Focus ‘21-’22 (and ongoing)

- We’ll evaluate and qualify the representative selection of profiles and pages, so that it is not only the most popular profiles that are harvested.

- Status:
  - We are constantly evaluating and qualifying our selection and we refine our selection using:
    - Our curiosity, research, daily stand up meetings / sharing knowledge
    - Scrape/crawl content to identify networked content
    - Input from researchers
    - Business intelligence solutions (maybe in the future) / ranking lists.

- Because it’s so difficult to crawl some services/sites, we need to make sure we don’t spend our time on a big well-maintained list of impossible-to-crawl sites.
<table>
<thead>
<tr>
<th>Social media / Crawl technology</th>
<th>Twitter</th>
<th>Facebook</th>
<th>YouTube</th>
<th>Reddit</th>
<th>Instagram / Picuki</th>
<th>TikTok</th>
<th>Tumblr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objects, content_type_norm (status code 200) ca. 2000 - September 2022. Mainly harvested with Heritrix</td>
<td>html (179,793,374) other (5,196,610) text (659,696) image (212,195) video (300) pdf (36) audio (8)</td>
<td>html (94,005,499) image (6,810,653) text (6,289,421) other (3,807,962) video (4,922) pdf (2,175) audio (487)</td>
<td>html (72,820,874) other (42,356,226) image (10,781,399) text (10,502,889) video (121,271) audio (5) pdf (1)</td>
<td>html (7,947,014) other (2,034,634) image (96,097) text (37,083)</td>
<td>html (4,596,213) other (251,032) image (14,148) text (141,148) pdf (192) / (8,601) pdf (2)</td>
<td>html (2,694,009) video (935,787) other (273,260) text (252,233) image (627)</td>
<td>image (33,949,656) html (5,188,885) text (976,429) other (690,536) video (14,427) pdf (19) audio (17)</td>
</tr>
<tr>
<td>Heritrix (issues with dynamic or embedded content - like YouTube videos)</td>
<td>Blocked</td>
<td>Captchas/log-in pages. Some luck with m.facebook.com (no comments)</td>
<td>Blocked - awaiting IP-whitelisting from Google</td>
<td>Using old reddit</td>
<td>Via picuki</td>
<td>Some content</td>
<td>10.000 + curated sites</td>
</tr>
<tr>
<td>Archivelt</td>
<td>Probably. Awaiting using API-solution</td>
<td>Lots of captchas/log-in pages. Sometimes content beyond &quot;more&quot;/scroll</td>
<td>Ad-hoc automated from a seedlist</td>
<td>Curated content</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>archiveweb.page</td>
<td>Manually with autopilot</td>
<td>Ad-hoc semi manually (posts &amp; comments). Autopilot (autoscrolling of content) works on some type of content.</td>
<td>Ad-hoc manually</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Browsertrix &quot;Cloud&quot; (Webrecorder / IIPC)</td>
<td></td>
<td>Hope to use this more for logged-in crawling</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>API/Twarc/Wget - Pilot Awaiting approval</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discovery/Playback in SolrWayback</td>
<td>API/Twarc/Wget/Pilot-playback. Earlier Heritrix crawls looks great</td>
<td>To some degree. Difficulty playing embedded videos. Need for text to be indexed</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>To some degree. Difficult to see videos in context</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Crawling currently
Not crawling currently or issues crawling
Probably able to crawl with technology

Status 2022 - Social media / crawl technology
Social Media (some of it) vs. the entire web archive (status code: 200)

Twitter, Facebook, YouTube, Reddit, Instagram, TikTok, Tumblr

Entire web archive

506,969,945
27,451,751,557
Objects (status code 200) pr. domain

Social media - objects
Objects (status code 200) pr. domain

Reddit
Instagram/Picuki
TikTok
Tumblr

Social media - objects
Some of the extra content we nominated and tried to crawl in our Corona event harvest.
A way forward / ideas

We need to focus on efforts that actually adds content to the archive. A list of nice-to-have profiles/sites is not enough

- Getting Twitter API harvest in production at scale

- Week 46 and other focused efforts
  - Automatic/semi-automatic, high fidelity, crawling with Browsertrix/Webrecorder technology. "Small" amounts of the most important sites/profiles. Facebook-profiles crawled once a year getting posts and comments a year back or more.

- Identifying valuable content on services/sites that can actually be crawled

- Playback and search needs work as well. One thing is to harvest eg. Facebook, another is to make search/discovery/playback work at our end.
**SolrWayback**

- Avanceret søgning og visning af arkivet
- Indhold tilgængeligt 1-2 måneder efter husholding
- Tools: Wordcloud from domain, Link Graph, Domain stats, HTML-tags, Link graph Graphi export, Nirgram, Visualize data, GPS image search, Harvest calendar, PWID, Website preview, View page resources

**OpenWayback**

- Søgning på specifik URL
- Indhold tilgængeligt 2 dage efter husholding
- Enkel kalendervisning
NO Collections

"Only” one big archive

- At the moment we don´t show specific collections but we do make event- and special harvests (harvests have an unique ID that can be used to show defined collections at a later point)

- Netarkivet is only available for researchers at Danish institutions (Ngram search is one exception) and users apply for access to a specific research project

- Instead of curated collections, our users can use free text search in SolrWayback and can apply for a data dump delivery for further use.
"sociale medier"

Type: html, webpage @ au.dk
Date: 31/10/2018

Highlighted content:
"Sociale Medier Blog om sociale medier Menu Skip to content Home Sociale Medier Forelæsninger og"

Images: showing 4 out of 11 See all images
Visualization of search result by domain
Twitter search

Domain: twitter.com AND "anders klindt"

Results:
Showing 0 - 17 of 17 entries matching query.

1. ANDERS KLINDT MYRVOLL (@ANDERSKLINDT) | TWITTER
Type: html, web page @ twitter.com
Date: 16/05/2020
Url: https://twitter.com/AndersKlindt

Highlighted content:
'@tolk09163 ...https://twitter.com/thewelkthatwas/status/1141677772160X10304 ... Anders Klindt Myrvoll'

Images: showing 4 out of 27

Anders Klindt Myrvoll
@AndersKlindt
Programme Manager at the Danish web archive, Digital Cultural Heritage. The Royal Danish Library

Tweets
88
Følgere
132
Følgere
89
Likes
199

Tweets & svar

Medier

Anders Klindt Myrvoll @AndersKlindt 31. mar.
På Det Kgl. Bibliotek og i Netarkivet knokler vi for at dokumentere Coronakrisen. Vi du være med? Se mere her: kb.dk/nyheder/fortæl... Kom også gerne med dit input til hæmmeder direkte her til byl@39v12523 (googlebom)

1 5 8

Anders Klindt Myrvoll @AndersKlindt 19. jan.
Spændende job som informationsspecialist/bibliotekar i København på min arbejdsplads. Frit d. 27 januar, @borgerbaman candkate Hv-

Ny på Twitter?
Tilmed dig nu for at få din egen personlige tidssæt!

© 2020 Twitter

Twitter Heritrix harvest
Twitter API harvest – search with specific type-facet. Pilot project (not in production yet)
Examples of playback - Twitter API Pilot project (not in production yet)
Examples of playback, including video - Twitter API Pilot project (not in production yet)
Welcome to Facebook!

Facebook is a social utility that connects you with the people around you.

Facebook is made up of lots of separate networks — things like schools, companies, and regions.

You can use Facebook to:

- Share information with people in your networks.
- See what's going on with your friends.
- Look up people around you.

Facebook © 2006
Det Kgl. Bibliotek
Royal Danish Library

Picture from P6 Beat Facebook-page – probably screenshot from [https://www.bbc.co.uk/programmes/p04ghl0n](https://www.bbc.co.uk/programmes/p04ghl0n) @ 01:46
Jeff Buckley performs Grace on The Late Show

Jeff Buckley performs the powerful title track to his album Grace on The Late Show in 1995. The singer-songwriter, who died in tragic circumstances two years later, would have celebrated his 50th birthday on 17 November 2016.

17 November 2016
© BBC

Featured in...

From the Vaults
BBC Arts
Our pick of treasures from the BBC archive

https://www.bbc.co.uk/programmes/p04ghl0n @ 01:46
Facebook - Playback in SolrWayback with comments – text not indexed for free text search – crawled 2019
Effektivt arbejde kræver nogen gange…
Effektivt arbejde kræver nogen gange minimalistisk techno som VCMG: https://open.spotify.com/album/0GObYf9FZ28MAz1NMjyt3UkK — Andre gange må man jo bare lave det selv!:) #Juno808HeadPeggarator:DtFenstingenTrommemaskine5v

**Anders Klindt Myrvoll**
15. marts 2018

... og så derefter gøre det med en lidt mere kompliceret maskine...

**De1**

... hvis du mangler noget at lave trommemusik med, så tjek maschine ud fra Native Instruments

**prompt666**

... er det sådan en idé...? Ellers kunne det jo godt blive lidt dyrere med sådan en karl her: https://www.youtube.com/watch?v=AMj0vQ5W5

**Old school House with Roland TR-808, Juno-60, SH-101, TR-809**

A Anders Klindt Myrvoll læskerbiskon 😘

7 kommentarer
Want to find out more?

- https://www.kb.dk/en/find-materials/collections/netarkivet
- Exploring special web archive collections related to COVID-19 - The case of Netarkivet
- An international review of Social Media Archiving initiatives
- SolrWayback 4.0 release! What’s it all about?
- SolrWayback 4.0 release! What’s it all about? Part 2
- https://github.com/netarchivesuite/solrwayback/
- Try SolrWayback Live demo made by National Széchényi Library, Hungary
- https://labs.statsbiblioteket.dk/linkgraph/
- https://labs.statsbiblioteket.dk/netarchive/ngram/#/

Contact Anders/KB: write to kb@kb.dk
https://www.linkedin.com/in/andersklindt/
https://twitter.com/AndersKlindt