Less is more
reduced broad crawls and augmented selective crawls

—a new approach to the legal deposit law

Sabine Schostag
Web curator

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
  - Addressed to a Danish audience

Translation of Act No. 1439 of 22 December 2004
Unauthorized version
Some facts on the Danish national web archive

- Founded as a coorporation of the two national libraries, the Royal Library and the State and University Library (now merged into the Royal Danish Library)
- Curator Tool: Netarchive Suite
- Crawler: Heritrix (2005-2016: Heritrix 1, 2016-> Heritrix 3)
- Size: 784 TB (2017-06-04)
- Read more: [http://netarkivet.dk/in-english/faq/](http://netarkivet.dk/in-english/faq/)

Overall crawl strategy 2005-2016

- 4 broad crawls (in depth going) in two steps each per year (first step with a limit of 10 MB per domain, second step with a limit of 2 GB per domain)
- Selective crawls from about 100 frequently updated sites (mostly News sites) with a variety of frequencies from 6 times a day to monthly and a depth from frontpages only to 4 levels.
- About three event crawls every year (events, that increase activities on the web, e.g. parliamentary elections or the terror attack in Copenhagen in 2015)
- Special crawls (YouTube crawls, one last crawl of a site, which is going to be closed down,…)
Still 4 broad crawls

- Goal: 1-2 full broad crawls in 2 steps, 2-3 broad crawls in one step (max. 100MB/domain)
- In practice 2016: 3 broad crawls in one step (max. 100MB/domain)

Storage: size of one full broad crawl

- 2016: ca. 40 TB
- 2009: ca. 9 TB
New strategy: More selective crawls

- Frequently updated sites (150-200 sites)
  - All national news sites (daily)
  - All regional and local news sites (daily/weekly)
  - Political players (politicians, commentators, ..., daily)
  - Peculiar and experimental sites
  - Sites reflecting the contemporary society
  - Representative profiles from Social Media (*special strategy in preparation*)

New strategy: More selective crawls

- Crawls, withdrawn from the broad crawls (aprox. 250 sites)
  - Ministries and government agencies (4 times/year in depth)
  - Ultra big sites (e.g. dr.dk, 4 times/year in depth)
  - YouTube videos (continuously)
New collection strategy

- 2016 ff

![Diagram showing coverage over time with different collection types]

What are the benefits? Less redundancy

- Analyse of crawl.logs of one broad crawl 2014: there is redundant content in the archive, because we crawled every domain in depth four times a year.
- Lots of websites are cumulative and changes are only made on the topmost levels.
What are the benefits? Fewer harvest definitions

- 2015/16: 65 harvest definitions

- 2016 - about 20 active harvest definitions
  - websites appear in only one harvest definition.
Other benefits

- Easier to extract corpora (relevant for giving access)
  - Today: access restricted to researchers and Master degree students (under existing Danish data protection law)
  - Tomorrow (vision): broader access, at least to parts of the archive (a corpus, e.g. public bodies’ websites)
  - A website/domain will only appear in jobs from one definition. Extractions from the archive can be done on job level.

Selective Harvest

<table>
<thead>
<tr>
<th>Domains</th>
<th>Class configuration</th>
<th>Resource Store Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>denmarkarchive</td>
<td>näturlærer og skole, view, intern</td>
<td>Danmarks-Nationalbibliotek</td>
</tr>
<tr>
<td>adk.dk</td>
<td>näturlærer og skole, view, intern</td>
<td>Danmarks-Nationalbibliotek</td>
</tr>
<tr>
<td>sitearchive</td>
<td>näturlærer og skole, view, intern</td>
<td>Danmarks-Nationalbibliotek</td>
</tr>
</tbody>
</table>

Less is more
reduced broad crawls and augmented selective crawls

—Thank you for listening

Sabine Schostag
sas@kb.dk
@saschossi