2017-10-03 Statusmeeting

Agenda for the joint KB, BNF, ONB and BNE NetarchiveSuite tele-conference 2017-10-03, 13:00-14:00.

Practical information

- Go to https://c.deic.dk/netarkivstyregruppe
- Login as guest
- Write your name
- Insert password: wayback

Participants

- BNF: Sara, Géraldine
- ONB: ~~Michaela~~ (unable to attend), Andreas
- KB/DK - Copenhagen: Stephen, Tue, Nicholas, Jonas, Birgit
- KB/DK - Aarhus: Colin, Sabine
- BNE: Mar
- KB/Sweden:

Upcoming NAS developments

- any started work on 5.4? https://sbforge.org/jira/projects/NAS/versions/12944

Status of the production sites

Netarkivet

We are preparing a 2-days workshop for Netarchive curators on harvesting social media. Hopefully the outcome will be usefull for our coming event harvest on local and regional elections on 21 November. We also aim to use BCWeb with external partners on the election event harvest.

The developers are going to have a workshop in the middle of October. The curator wishes are as follows (in order of priority):

- Replay of https-pages in Wayback
- Improvement of Heritrix and integration of supplementary collection tools (e.g. brozzler)
- Introduction of a (technical) collection concept. This will give us the ability to integrate data collected before and without NAS.
- Improvement of Access
- More automated QA

Most likely we wil not be able to perform a full broad crawl with 2 steps this year (our last full broad crawl is from the beginning of 2016), because of our problems with Heritrix 3 Remote Access. We expect to be able to solve this problem with NAS 5.4, which will be implemented after having finished the compression of the archive in the beginning of 2018.

Since January 2017 we only harvested about 25 TB

In the beginning of September 2017 Netarchive was blocked by about 54.000 domains (out of 1.32 Mill. Domains)

The implementation of “Web Danica” (automated identification of Danish web content outside .dk) is ongoing.

The migration of documentation from the old “MediaWiki” to Jira is finished.

BnF

There have been several changes in the team over the summer. Pascal Tanésie has arrived as assistant head of the digital legal deposit team, and Vladimir Tybin has joined the team as digital curator. Sophie Derrot has left the BnF to take up a post at the Institut national d'histoire de l'art.

Our second test broad crawl, with the complete seed list, is nearly finished. The amount of data crawled in this test has proved to be higher than our budget estimates, mainly because there is no deduplication for this first broad crawl with H3. We will analyze the figures in detail and adapt the budget accordingly.

We are also using our new infrastructure for the tests: the crawlers are more powerful and faster but they use more bandwith. We will therefore need to reduce the number of crawlers from 40 to 35. We had set the duration of each job to 3 days but this has proved to be too much, for the real crawl it will be betwen 2 and 2.5 days.

This week we aim to transfer all our crawls onto the new infrastructure and the next week the real broad crawl will start.

ONB

- At the moment our 5th broad crawl is running
- We are already started our selective crawl for the parliamentary election on October 15th.

BNE

Our IT team has been working on the implementation of NAS 5 and they installed a complete preproduction environment (connected to CWeb) of NAS 5.3. They run several tests and checked that some problems we had been experienced with NAS 4 (especially related to security certificates) have been solved with NAS 5.

We expect to have a complete production installation of NAS 5.3 by the week of October 23rd. Once this version is installed in a production environment, our first task will be to run a domain crawl of .gal domain (the domain attached to Galicia). We expect to have it finished in about 3 days.

We've been also concentrated in curating our Catalan Politics collection, which was a thematic collection, but it's indeed a mixture of thematic and event collection. We decided to keep it as it previously was (a thematic collection), but adding new seeds, launching it more frequently and tuning some configurations.

We finally made access available to our web archive at the beginning of July. The online access only allows seeing what captures we have from every site, but the archived content itself is only accessible in our premises and the ones at the regional libraries with legal deposit competencies. Some of them have opened also this access to their users. It is not allowed to download or copy any part of the web archive, due to our copyright law limitations.

We are also preparing our annual workshop with regional web curators, scheduled for November 20th, to review the state of the art of our collaborative project of web archiving and non-print legal deposit.

Next meetings

- November 7th

- December 5th
- January 9th, 2018

Any other business?