2017-11-07 Statusmeeting

Agenda for the joint KB, BNF, ONB and BNE NetarchiveSuite tele-conference 2017-11-07, 13:00-14:00.

Participants

- BNF: Sara
- ONB: Michaela, Andreas (need to leave at 13:45)
- KB/DK - Copenhagen: Stephen, Tue, Nicholas
- KB/DK - Aarhus: Colin, Sabine
- BNE: -
- KB/Sweden: -

Upcoming NAS developments

- ongoing work on 5.4: https://sbforge.org/jira/projects/NAS/versions/12944

Status of the production sites

Netarkivet

- Our third broad crawl ran from September 13 to September 25 – with a budget of 10 MB per domain, that is to say we ran our usual step 1. Due to unsolved problems with H3 we will not be able to run step 2 (budget normally 100 MB)
- We are preparing the event crawl of the local and regional elections on November 21. As our selective crawls cover the news media part of the elections, we will exclude them from the event crawl. We talked about using the last broad crawl for 2017 as a “back up” for the event crawl by starting it just after the election day. But as we won’t be able to run an “in depth going” broad crawl before spring 2018, so this will be no option. Anyway, focus will be on Social Media (Twitter, Facebook, YouTube) NGO’s, companies, other stakeholders.
- We hope to get hints and help from the to days Social Media workshop. The first day was very fruitfull. It focused on how to identify relevant profiles, content etc on Twitter and Facebook. The second day will be about capturing content (API’s etc.). After the second day (on Monday) I’ll provide any information that would be useful for you.
- We implemented BCWeb in our production system, our intention is to use it for the election event crawl. But there are still some open questions, especially the transfer in connection to our new way to build configurations (which do not include hops) is a big issue to be solved.
- We started testing BNF’s NAS preload tool for the activation/deactivation of domains and cleaning up of their seeds concerning the broad crawls.
- Our Webdanica project (automatic finding Danish content from TLD’s other than .dk by capturing outlinks from domains archived in Netarchive) is almost ready for going into production.

BnF

- Our 2017 broad crawl was launched on the 16th October. The settings are 1500 URLs per domain, with a limit of 3 days per job. Our prediction of the overall volume based on our tests seems to have been underestimated: we had calculated around 77 TB with these settings and after three weeks of crawling we are now expecting a final volume of around 97 TB. This is still within our overall storage budget but we are keeping a close watch on the volume of data collected. So far we have encountered no major problems, both H3 and the new infrastructure are functioning correctly.
- We are also continuing to work on updating our full-text indexing process with the aim of indexing our news crawls since 2016. We have been updating the indexing schema to follow recent developments on warc-indexer and we will be working on the organisation of the index to improve query performance. The research project that will use this index to study neologisms is starting this week, so we will be working closely with a research engineer over the next few weeks.
- We are working on BCweb to integrate KB developments in the 5.3 release and fixing some minor layout and redirection bugs.

ONB

- Our Domain Crawl for this year just finshed a few days ago (With Nas 5.3 and all the expected problems - a lot of times we had to terminate jobs by calling the kill script, a couple of times we had to stop NAS and had to clear the message queue, due to too many messages, which caused inactivity of NAS)
- Now we are doing some postprocessing work (indexing, reporting)
- Next step is a redeployment of 5.3.1 to reproduce our problems we had in summer and to save the log files for further discussion. If that fails again we go back to 5.3. which works for selective crawls only without a problem
- We have just finished our domain crawl. We started it in August with a 10 MB limit and archived approx. 5 TB of data. We will not do a 2nd stage due to our storage budget. We will analyse the outcome, because the behaviour and storage consumption has changed with H3. (In 2015 we used 0,5 TB for stage 1 with 10 MB.)
- Our domain crawl interval will change to annual crawls. Unfortunately, we will not get more storage. It will change from 2 TB/10 TB to an annual budget of 6 TB.
- Parliamentary elections took place mid-October and our crawl continues until a have a new government.
- In November we start another crawl of our women/gender collection.

BNE

Next meetings

- December 5th
- January 9th, 2018

Any other business?