Update on NAS latest tests and developments
Review of homebrewed heritrix extensions and feedback on tests.

Status of the production sites
Netarkivet
We are preparing the start of our next broad crawl Mid-March. We are awaiting Heritrix IIPC-release with the last of our pull requests.

The library’s executive board wants to keep only one online copy of Netarkivet due to costs for online copies (at the moment we have two, one in Aarhus and one in Copenhagen).

A steering committee, project group and IT-developers are working on whether and how we can achieve one online platform for Netarkivet with the same services and performance – and still with the right level of security in terms of preservation.

A shitstorm provoked by a commercial for SAS (Scandinavian Airlines - https://youtu.be/ShfsBPvNcTg) claiming that everything typical Scandinavian was stolen from other nationalities, showed, that we have to react very quickly. We run a kind of mini event harvest focusing on the reactions on Twitter and Facebook.

We decided to welcome a French student for about three month in autumn. We will probably invite him to a video conference soon to talk about the tasks he could work on for Netarkivet.

BnF

ONB

BNE

We celebrated the 10th anniversary of the Spanish Web Archive. We organized a conference where we had the opportunity to share experiences with other colleagues.

New collections:

- We are working now in two event crawls about the elections in two regions: the Basque Country and Galicia
- We want to launch a new collection about feminism this month

Serials broad crawl: We were preparing a list of serials urls in free access. They are almost 8,000 and we launch a kind of broad crawl to harvest them.

KB-Sweden

(At last we write something here. We apologize for not doing it before.)

Recap: We have run selective harvests successfully since middle of 2018 but had during 2019 lots of problems with running NAS snapshots above a certain size level (number of domains). There was a bottleneck in the system but it was hard do figure out what. At last we found it: the Postresql database server was overwhelmed and database requests were queued up, which made the system slow and hard to monitor (as the GUI updates was out of phase).

Eventually we realized what was the bottleneck, added some indexes to the database and suddenly everything went like clockwork! At least technically.

So in December we could complete the first part of our broad crawl (just 500 kByte limit). And in January we started part 2, with around 500.000 domains remaining and limits 2 Gbyte and 50000 objects. It has now run over 90 % of the jobs, so will probably be done within this week. Very good!

Things we discovered when monitoring is the large amount of sites which are same kind of shop, displaying many thousands of products, with a couple of images of each. And sites related to sport activities, having tons of match results and player statistics. This, combined with errors in links creating looping URL:s can lead to millions of URL:s in queue. It takes some time before suchs jobs reach the 50.000 objects limit, so we have been monitoring and deleting URL:s in the queue now and then.

Next meetings

- April 7, 2020
- May 5, 2020
- June 9, 2020
- July 7, 2020
- September 8, 2020
- October 6, 2020
- November 3, 2020
- December 8, 2020
- January 5, 2021

Any other business?