2022-03-08 Statusmeeting

- Update on NAS latest tests and developments
- Status of the production sites
- Next meetings
- Any other business?

Agenda for the joint NetarchiveSuite tele-conference 2022-03-08, 13:00-14:00.

Participants

- BNF: Clara, Auriane
- ONB: Andreas
- KB/DK - Copenhagen: Anders, Thomas, Stephen, Tue
- KB/DK - Aarhus: Colin
- BNE: Alicia, Miguel, José
- KB/Sweden: Peter, Par, Jonas

Update on NAS latest tests and developments

There has been little significant work on NetarchiveSuite since the Release of 7.3 at the end of January

- Some minor improvements to the handling of hdfs-cached warcfiles for hadoop mass-processing
- Some speculative work on making deduplication-indexing more concurrent - shelved as it is not currently a priority

An important question we have to deal with is how to manage the fact that the Netarkivet configuration of NetarchiveSuite has now diverged very markedly from that used by most other users. In particular we no longer use the ArcRepository application (The ArcRepository interface remains in use), BitarchiveServer, BitarchiveMonitorServer, or ChecksumFileServer. According to the usage page at Institutional Usage of NetarchiveSuite this is a particularly strong divergence from the OnB setup - but we are missing data from BNE and KB/Sweden.

So what is the future of these components? I think that we will always need to offer a fully-functional Quickstart environment so we will still need to be able to store files and run batch-jobs. But that can be done with local files (as at BnF) and doesn’t require any kind of distributed repository. We don’t need to remove any code, but in the long run I don’t think KB/Denmark can assume responsibility for maintenance of those parts of the codebase we don’t use ourselves, so that the distributed ArcRepository and associated components would ultimately either have to be provided only “as is” or maintained by the institutions that continue to use them.

Status of the production sites

Netarkivet

- Broad crawl
  - Step 1 finished in great fashion with new Bitmagasin
  - Step 2 will start March 7
- War in Ukraine event harvest ongoing. Helped DCN-institutions from Ukraine with info/best practice
- SolR/Wayback
  - Using the ongoing harvests (SolR-index/shard) as QA-platform for curators - Q2 2022
  - New Shard built
- Focus on paywall content and IP-validation to get the most data possible
- Twitter API-harvest - still pilot project, but very relevant with lots of activity regarding Ukraine

BnF

First, the Winter Olympic Games harvest ended at the end of February and a new crawl dedicated to the Winter Paralympics has been launched last week. About 14 million URLs were collected in February, including almost 1.4 million Twitter URLs for a total of 0.57TB.

We also decided to make another attempt to collect Instagram in 2022. After several tries, we succeeded. 73 Instagram accounts on the theme of the Olympics were collected, that is to say about 7 000 URLs.

Finally, last December we opened a participation form until the end of January so that the public could indicate sites to be added to the Intelligence Artificial harvest.

We received 23 e-mails including about 60 suggestions of websites to crawl.

ONB

BNE
We will run our third annual broad crawl of e-serials in open access next month. This is the third year in a row that we launched it. We obtain the url list from our catalogue where all the serials that request ISSN are catalogued. Every year we enhance the list of urls thanks to the quality control we carry out after each crawl. We will harvest about 9,000 websites this year.

The Library is involved in a renovation of its technological structure, so we are studying when we will be able to carry out the annual broad crawl. It will probably take place before summer. The renovation involves the acquisition of new servers and solves some problems we had last year with the storage arrays. Once the new technological structure has stabilized our next step is the NAS upgrade, that we think we can tackle next May. We also want to start to index the current crawls in SolrWayback and we will address later the indexing of everything harvested prior to 2022.

KB-Sweden

Next meetings

- April 12th
- May 10th
- June 7th
- July 5th
- September 6th
- October 4th
- November 8th
- December 6th
- January 10th, 2023

Any other business?