2023-12-05 Statusmeeting

- Lastest NAS evolutions
- Status of the production sites
- Next meetings
- Any other business?

Agenda for the joint NetarchiveSuite tele-conference 2023-12-05, 13:00-14:00.

Participants

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

Lastest NAS evolutions

Lastest release NetarchiveSuite 7.5: https://sbforge.org/display/NAS/NetarchiveSuite+7.x+Release+Notes

Status of the production sites

Netarkivet

- 4th Broadcrawl 2023- step 2 - all jobs running - expected to finish before christmas holidays
- Still testing on site installation of Browsertrix Cloud as well as the beta-installation (Webrecorders installation) - will intensify the coming month - possibility of customnaming of WARC-files is a must
  - Lots of great possibilities
- PyWb and CDX-indexing done.
  - PyWb instance on SolrWayback Stage (alternative playback).
  - Awaiting to move CDX-index to CPH-servers to enhance performance. The index needs to be close to the WARC-files and not between 2 firewalls.
  - Crawling and playback of advanced sites like Instagram, Facebook is still an issue. We thought our PyWb-installation would playback Instagram well, we made some tests earlier that indicated that, but seems there’s something wrong in PyWb playback with OutBack CDX (using JWARC as converter from WARC to CDX). We might talk to the community about these issues. All relevant in terms of PyWb roadmap, differences between PyWb and Replayweb.page (Browsertrix Cloud) and more. It would be great to have the same playback in PyWb as Replaywebpage (native or in Browsertrix Cloud)
- Finished week 46 special crawls of Radio-TV websites.
- New Citrix/VLan for Netarkivet ready to go into production (will be able to replay aprox. 70 mio. Flashsites)
- Data delivery. At the moment we have to exclude many sites were content can be obtained elsewhere as a service

BnF

Our 2023 broad crawl is about to end. It will have lasted 7 weeks, so one week more than last year due to the reduced number of crawlers used. The budget should reach 150 TB.

The Skyblog harvest is officially over and lasted 50 days. We crawled more than 1.8 billion URLs (12 millions blogs) and nearly 37 TB of data. The Orange personal pages harvest has been launched and should finish at the end of the week. It will be a smaller collection of 1.8 TB.

Finally, we have just launched the "Social movements" and "Solidarity" harvests, for which the projected budget is 1.95 and 1 TB respectively. The Solidarity harvest has been expanded with new selections around the theme of disability: association sites, employment incentives, forums, sports for the disabled.

ONB

BNE

KB-Sweden
Next meetings

- January 9th 2024

Any other business?