2022-10-04 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 2022-10-04, 13:00-14:00.

Participants

- BNF: Auriane, Clara
- ONB: Andreas
- KB/DK - Copenhagen: Anders, Thomas (vacation), Stephen (maternity leave), Tue
- KB/DK - Aarhus: Colin
- BNE: Alicia, Miguel, José
- KB/Sweden: Peter, Pär, Jonas

Update on NAS latest tests and developments

NAS 7.4.3 is released. This fixes two bugs. NetarchiveSuite 7.x Release Notes

Status of the production sites

Netarkivet

- 3rd broad crawl ’22 almost finished. We are aiming for ending it october 10th, so we can have the 4th broadcrawl for 2022 (which is the norm)
- Focus on Paywall and IP-validation have payed off. We get important content from quite a few more sites now.
- Anders attended:
  - Wanted: Social Media Data-conference in Brussels: https://www.kbr.be/en/agenda/wanted-social-media-data/ with the presentation Social media archiving at the Royal Danish Library_Sep 22.pptx
  - Digital models in humanities research - https://www.it-vest.dk/events/conference-about-digital-models-in-humanities-research - pretty interesting and a lot of interest in web and social media data
- Almost finished with the updated JWAT for validation of Warc-files
- Our accesplatform - SolrWayback - is getting it’s own Citrix vLan and will have flash player, as well as Gephi and R installed - so will be a small workspace.
- Solrwayback: Tranferring internal issues in Jira to https://github.com/netarchivesuite/solrwayback. Also great to see more institutions using Solrwayback and contributing to the code (bugfixes and hopefully more in the future)
- Lots of data dump deliveries this month and in the horizon.
- CDX-summary of Netarkivet’s holdings. We are not able to participate at the moment.
  - https://github.com/ymaurer/cdx-summarize

BnF

Our 2022 broad crawl is going to be launched. This year, it has been possible to increase the budget to 2700 URLs per domain, for a total around 145 TB. Each job will end 3 days after its launch. The crawl is expected to finish in the middle of November.

We have started the preparation of two virtual guided tours. The first one will be published in December 2022 and will highlight our collections relating to artificial intelligence. The preparation of the second one has just started. It will concern Elections collections (2015-2022) and the publication is scheduled for the first quarter of 2023.

In October, we will start working on our next internal harvesting workshop scheduled for November. It will be devoted to the harvest of Podcasts and sound documents and we will work with the Sound, Video and Multimedia department of the BnF.

ONB

BNE

This month we are focused on the preparation of the workshop about non-print legal deposit and web archiving within the framework of the legal deposit working group of ABINIA (Association of Ibero-American States for the Development of National Libraries of Ibero-America). It will be held on October 5th by Zoom, with the confirmed attendance of 13 countries.

We are working to solve the problems we have when we want to harvest Twitter. We have created a new template of 2,000 objects that reduces the saving time and avoids 429 errors. Now we are testing in Pre.
Next meetings

- November 8th
- December 6th
- January 10th, 2023

Any other business?