2022-09-06 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 2022-09-06, 13:00-14:00.

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

Update on NAS latest tests and developments

Release of NetarchiveSuite 7.4.2

Status of the production sites

Netarkivet

- Broad crawl, Step 2 of the 3rd broad crawl 2022 started
- Heritrix: Using RSS and extractor modules to get more content. Eg: issuu-PDF's
- Focus on Paywall and IP- validation. Some News-sites responds quickly others not at all.
- Preparing for the Besocial 2022 conference
- SolrWayback "live"- QA still up and running and is great for QA (although some resources are missing due to deduplication).
- IIPC Browserbased-crawling project - workshop today at 16-17 CET
- Progress is made on the updated JWAT for validation of Warc-files

BnF

First of all, we welcome Nola N'Diaye in our team as harvesting manager and assistant head of the digital legal deposit team. She succeeds Pascal Tanésie who is retiring in December.

Last month, nas-preload version 9.1 and NetarchiveSuite version 7.4.1 have been released. The new version of NAS includes several improvements and evolutions which will be useful for monitoring the crawls: display of the compressed data size of the WARC files produced by each running job, distinction of the queues types on Progression and Queues page, bug fix on the possibility to use a regex with a backslash on Browse/Delete frontier...

We are also going to launch a test broad crawl this week. Our production crawl will be launched in October.

The crawl stemmed from the LIFRANUM project which concerns digital French-speaking literature websites, ended last week. 1089 seeds (websites, blogs hosted on several platforms such as wordpress.com, over-blog.com, etc...) have been harvested. We also crawled separately a few thousand contextual contents webpages with a dedicated job. The selection step was made with Hyphe, a web corpus curation tool based on a web crawler.

Finally the IIPC webinar "Web Archiving the War in Ukraine" took place last Wednesday. On this occasion our colleagues Vladimir Tybin and Anaïs Crinière-Boizet presented, with Kees Teszelszky, the "War in Ukraine" IIPC collaborative collection led by the BnF and the National Library of The Netherlands.

ONB

BNE

This month, we are working on the organization of an online workshop for countries are part of ABINIA (Association of Iberoamerican States for the Development of National Libraries of Iberoamerica) It will be in October or November. We want to show how the Spanish web archive works through its collections, infrastructure and operation. Many Lationamerican countries are beginning to consider the creation of their web archives and we want to help them in their first steps.

Lately. We have had some problems to harvest Twitter. Some days we have errors 429 in NAS report, we think it is for high number of account that we are collecting, currently about 3,000, most of ones weekly, we try to reduce the number to avoid this problem.

National Library of Peru is interested to use NAS. Would it be possible to invite them to come to next NAS meeting?
KB-Sweden

A broad crawl of the .se domain is under way since before the summer. Estimated to be finished in mid October. The admin servers and half the harvesters are running version 7.4.2 and the remaining harvesters are running 7.4.1 until they are done with their current job. Updating on the fly is working rather well with some shellscript wrappers.

We have experienced some problems with misconfigured object limits.

Next meetings

- October 4th
- November 8th
- December 6th
- January 10th, 2023

Any other business?