2022-06-07 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 2022-06-07, 13:00-14:00.

Participants

- BNF: Sara, Clara, Auriane
- 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

Status of the production sites

Netarkivet

- Step 2 of the first broad crawl will start 08 Jun 2022
  - Focus on cleaning up regular expressions
  - Limits of domains
  - New harvester-server up and running soon
- Outsourcing of harvesting - status: will be kept in house.
- SolrWayback "live"-QA up and running
- RSS-Heritrix module tested and still needs some focus
- IIPC Browserbased-crawling project is proceeding. We have an update meeting tonight and have had input during the IIPC GA-sessions.
- A lot of data delivery for researchers at the moment.
- Working on updated JWAT for validation of Warc-files (communication with Nicholas and estimating budget/finding a way to this now internally at KB)
  - Support sha256
  - Missing support for modern gzip os
  - support for [[ ]] in urls
  - http request headers
  - warc 1.1 support

BnF

Last week, we had a meeting to prepare the program of the 2022 broad crawl. In this context, an overhaul of nas-preload and developments concerning NAS are planned. The registrars have been contacted and we’ve already got most of the lists. Two new TLDs from overseas departments and territories have been obtained. The launch of the broad crawl is scheduled for October.

The Official Publications harvest has been launched last week and will last at the end of June. This harvest includes websites of ministries, public establishments, independent administrative authorities and local authorities. Nearly 900 websites have been selected.

Finally, our next Videos harvest is in preparation. We are encountering some difficulties because we have changed the metadata extraction tool. The number of metadata extracted and therefore videos to download is indeed much greater than with the previous tool, which raises budget issues.

ONB

BNE
The broad crawl 2022 of the .es domain ended on May 19th. It has taken 21 days (compared to 25 days last year) with a limit of 150 MB per domain and 71 crawlers. This year the harvest was carried out through the BNE internet line. This has meant a reduction in the number of days we have used. In terms of results, we crawled 69 TB. In terms of documents harvested, we saved 3.54% less. This may be due to the fact that we have eliminated earlier the jobs that were stuck due to poor site configuration. If we combine both factors (fewer but larger items) we assume that we have a higher quality collection.

The broad crawl of journals was completed in April. The number of websites collected with electronic serials was more than 12,000, that is around 3.4 terabytes.

KB-Sweden

Next meetings

- July 5th
- September 6th
- October 4th
- November 8th
- December 6th
- January 10th, 2023

Any other business?