2022-05-10 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 2022-05-10, 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

The User Documentation for NAS 7.3 has finally been released.

The latest snapshot includes a fix to CrawlRSS which is currently being tested.

(Issues mentioned: Memory issue with regex in v. large crawl logs. Harvesthistory is slow but Tue has a fix.)

Status of the production sites

Netarkivet

- The first broad crawl will end this or next week
- Minor event crawl on EU/defence-election
- Flickr. Big effort to get Danish content before changes
- SolrWayback:
  - For the next maintenance sprint we are making a SolrWayback instance were web curators can access the SolrWayback-SHARD while it is being built rather than wait 1-3 months (depending on the amount of data that is harvested) for improved quality assurance
  - Anders presenting at the IIPC WAC 2022: #2: SolrWayback at the Royal Danish Library - Key Findings, Experiences and Future Aspects as part of - Full-text Search for Web Archives - https://netpreserve.org/ga2022/wac/abstracts/#Session_1
- RSS-Heritrix module ready for testing

BnF

The first part of our 2022 Elections harvest, launched last January, is about to end. The second part of the crawl concerns the legislative elections and will begin on May, 16. On this occasion, 16 public libraries all over France are also taking part in the selection process. Moreover our next Instagram harvest that is going to be launched this week is also dedicated to the 2022 presidential elections. Nearly 60 accounts will be crawled.

Finally, to celebrate 20 years of harvests about French elections, we have put online a new homepage of the "Archives de l'internet".

This week we are also going to launch our first TikTok crawl. This new harvest follows an internal workshop that took place last March. This crawl will also be on the theme of the presidential elections in France. We plan to collect about 50 accounts and 50 tags.

At the end of April, a new virtual guided tour has been published on the theme of biography on the web. It proposes to discover through 9 themes the different forms that biography can take on the internet.

The annual harvest is still ongoing and will continue until the middle of May.

ONB

BNE

- Last month, we launched the massive crawl of periodicals in free access for 3rd consecutive year . We have harvested more than 12.000 websites that hosted electronic serials.
- We have worked with our regional collaborators on a new event collection for the regional election in Andalusia.
- At the beginning of May, we have started our .es domain annual broad crawl, arround 2.000.000 domains and it is going to end in 25 days
KB-Sweden

About to start a broad crawl.

Preparing for selective crawls in connection with the general elections in September, plus video harvesting with youtube-dl and Twitter harvesting with Social Feed Manager.

About to replace the old tape-based wayback solution with Pywb and WARC:s collected with NAS.

Next meetings

• June 7th
• July 5th
• September 6th
• October 4th
• November 8th
• December 6th
• January 10th, 2023

Any other business?