2023-04-11 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 2023-04-11, 13:00-14:00.

Participants

- BNF: Sara
- ONB: Andreas
- KB/DK - Copenhagen: Anders, Thomas, Stephen
- KB/DK - Aarhus: Colin
- BNE: José, Miguel
- KB/Sweden: Peter, Pär

Update on NAS latest tests and developments

Status of the production sites

Netarkivet

- Second Broad Crawl will start soon
- Data dump of all text from Netarkivet to research project on making a new Danish language model in the works. see more here: https://github.com/kb-dk/kb-scripts/tree/master/all-text
- Awaiting invitation from Norways Nettarkivet to learn more about their archive.
- Twitter API! Still awaiting new solution. Considering contacting them.
- Focus on IIPC WAC 2023. Presentations uploaded and available. SESSION 8: BROWSER-BASED CRAWLING (password)
- Asked for PyWb-analysis to be prioritized for maintenance sprint (May)

BnF

Our internal harvesting workshop about Browsertrix finished at the end of March. A total of 10 testers participated and more than 80 crawls have been launched for 40 use cases analysed. Each tester completed a use case analysis grid in order to structure the test feedback. Our feedback will be summarised and presented to the community soon.

Within the framework of our internal project to improve our harvests, we are currently running tests on Twitter accounts in order to improve the harvest. All the selected accounts are not covered homogeneously by the harvest. Many images are notably missing. According to our tests, it might come from the mass of data that we try to harvest.

The Environmental issues and Artificial Intelligence harvests have been launched at the end of March and concerns more than 700 and 650 selections respectively. The AI harvest has been enriched by selections about prompt art and generative AI.

Finally, the international ResPaDon symposium entitled “The web: source and archive” was held in Lille from 3 to 5 April. It gave rise to many exchanges between researchers and library professionals around web archives.

ONB

BNE

Creation of a new event collection about the regional and local elections in Spain. In total 12 regions have elections and the whole country has local elections. We coordinate with the different web curators the seed selection and quality control. The elections are going to take place on May 28th.

The preparation of the broad crawl of open access journal has been finished. We will be launch it at the end of April.

We continue with the problems with Twitter. Tests under similar conditions give very different results and we don’t know why. Thanks to the BNF and especially to Clara for her help with the templates and these problems. We expect to find a solution soon, this year there going to be regional and national elections, and Social Networks are very important for us.
KB-Sweden

Jonas Linde has finished his 1.5 years as a consultant with the library, but we are already planning to use him further, both for shorter jobs and possibly for a longer period in 2024.

The second layer of our first broad crawl of 2023 has finished and we will start the third layer as soon as possible. We are aiming for at least one more broad crawl this year.

We now have our whole web archive (1997--) available for our visitors through Pywb. We are not yet permitted to use free text search (i.e. SolrWayback) because of limitation in the legislation.

We are moving parts of our e-legal deposit work from the methods we normally use (mainly RSS-based harvesting or web uploading) to web archiving with NAS. It has been difficult for government agencies and local authorities to identify what material is covered by the legislation.

Next meetings
- May 9th (cancelled)
- June 6th
- July 4th
- September 5th
- October 3rd
- November 7th
- December 5th
- January 9th 2024

Any other business?