2023-01-10 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 2022-01-10, 13:00-14:00.

Participants

- BNF: Auriane, Sara, Clara
- ONB: Andreas
- KB/DK - Copenhagen: Anders, Thomas (ill), Stephen (ill), Tue (day off)
- KB/DK - Aarhus: Colin
- BNE:
- KB/Sweden: Peter, Pär, Jonas

Update on NAS latest tests and developments

Status of the production sites

Netarkivet

- Broad crawl
  - 4th broadcrawl step 2 - 2022 still running - closing soon.
- Special harvest "Mystikkens Univers".
  - YouTube Using Archiveweb - 20,000 videos ~170GB!
  - Instagram using archvieweb.page and autoplay - gets slides+reels etc. Works ok, but only gets to 100 posts.
- Browserbased crawling for all IIPC-project awaiting funding for development this year
- Data from Internet Archive received
  - DK TLD-IA-crawls from 1996-2000 ~70GB
  - The dataset consists of all hosts in the Danish language in tab-separated format, sorted by SURT: host [tab] timestamp. The timestamp is the latest timestamp when the host was annotated and identified as Danish. It’s 9 MB small and consists of 680,838 sites.
- CDX-summary for research use approved and needs development work to be delivered for Aarhus University.

BnF

ONB

BNE

Lately, we have worked to create a new tool that will be able to search on the seed harvested by collections, topics, keywords and title. This tool extracts the information directly from CWeb and connects with OpenWayback to offer the calendar of harvests when people are outside of the library, and the calendar and web sites captures in the National Library facilities. Now, it is in preproduction, we plan to launch it in February or March.

We have made several tests with SolrWayback on our machines. We have indexed the broad crawl of magazines realized in 2022 (1.3 TB and 15,000 WARC). We have tested SolrWayback in different stress tests with a small index (140,000 documents), step by step, we are going to test with a bigger index of documents, so far, we have had good results. We have modified the indexation script to add a new functionality that allows us to choose a part of a WARC of a full set of WARCs for indexing a whole collection, not only a part.

We are studying how to config the SolrCloud with our technology infrastructure.

KB-Sweden

Next meetings
• February 7th
• March 7th
• April 11th
• May 9th
• June 6th
• July 4th
• September 5th
• October 3rd
• November 7th
• December 5th
• January 9th 2024

Any other business?