2022-12-06 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 2022-12-06, 13:00-14:00.

Participants

- BNF: Auriane, Sara
- ONB: Andreas
- KB/DK - Copenhagen: Anders, Thomas, Stephen, Tue
- 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 started a few weeks ago. More than 100 harvesters used concurrently (120 harvester capacity, 77 broadcrawlers)
  - Also working on other part of the broadcrawl with selective harvesters.
  - Bytelimit downgraded 61K shops to 10K maxobjects and 499MB maxbyte
- Event harvest
  - General election still running but will end soon
  - World Championship Soccer in Quatar - needs more seeds and then to be ended
- IIPC WAC 2023
  - 4 proposals approved:
    - SolrWayback: Best practice, community usage and engagement
    - Run your own full stack SolrWayback
    - Browser-Based Crawling For All: Getting Started with Browsertrix Cloud
    - Browser-Based Crawling For All: The Story So Far

- JWAT for validation of Warc-files updated - there might be some more work on documentation.

- Browserbased crawling for all IIPC-project proceeding. UX update will come soon with enhancements of exclusions and also using more explanations for each step/input.

  December update for Browsertrix Cloud:
  https://docs.google.com/document/d/1G7cy8mebn8mDA5w1rFn0mQ4DEN_B3Aite3C28ChneEs/edit#

  IIPC Just launched our new docs for browsertrix cloud at: http://docs.browsertrix.cloud/

BnF

- Our 2022 broad crawl ended on November 22nd. The harvest lasted around six weeks, that is to say one more week than last year for a budget of 2700 URLs per domain (instead of 2100 URLs in 2021). 3 billion URLs were crawled for a total of 151 TB.
- Next week, we are going to launch the "Social movements" and "Solidarity" harvests. 1037 and 473 websites are selected respectively. The harvests will last two weeks for a provisional budget of 1 TB for each.
- Our internal harvesting workshop dedicated to podcasts began in November and will end on December 16th. We studied several podcast platforms like SoundCloud, Ausha and podCloud.
- On November 25th, a webinar took place around the harvests and scientific practices of the electoral web, to bring the 20 years of elections harvests into relief. It was organized within the framework of the ResPaDon project, which aims to set up a network about web archives. The contributions made it possible to obtain feedback from the librarians who take part in the selection process and the researchers working on the subject.

ONB
BNE

KB-Sweden

Next meetings
  • January 10th, 2023

Any other business?