2022-11-08 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 2022-11-08, 13:00-14:00.

Participants

- BNF: Auriane, Sara
- 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
• Broad crawl
  • 3rd broad crawl ´22 finished end October (2-3 weeks more than anticipated)
  • 4th broadcrawl for 2022 started Nov 1st. (4 broad crawls is the norm)
  • We expect around 110TB, data for 2022.

• Event harvest on the General election including TikTok content using both Heritrix and archiveweb.page. Still running but will end soon

• IIPC WAC 2023
  • 4 proposals submitted

  | Submission Type / Conference Track: IN PERSON: 60 minute panel |
  | SolrWayback: Best practice, community usage and engagement |
  | Egense, Thomas (1); Toth, Laszlo (2); Eldakar, Youssef (3); Aubry, Sara (4); Klindt Myrvoll, Anders (1) |
  | Organization(s): 1: Royal Danish Library (KB); 2: National Library of Luxembourg (BnL); 3: Bibliotheca Alexandrina (BA); 4: National Library of France (BnF) |

  | Submission Type / Conference Track: IN PERSON: 60, 90, or 120-minute conference-themed workshop |
  | Run your own full stack SolrWayback |
  | Egense, Thomas; Eskildsen, Toke; Thøgersen, Jørn; Klindt Myrvoll, Anders |
  | Organization(s): Royal Danish Library, Denmark |

  | Submission Type / Conference Track: IN PERSON: 60, 90, or 120-minute conference-themed workshop |
  | Browser-Based Crawling For All: Getting Started with Browsertrix Cloud |
  | Jackson, Andrew N. (1); Klindt Myrvoll, Anders (2); Kreymer, Ilya (3) |
  | Organization(s): 1: The British Library, United Kingdom; 2: Royal Danish Library; 3: Webrecorder |

  | Submission Type / Conference Track: ONLINE: 45 minute panel |
  | Browser-Based Crawling For All: The Story So Far |
  | Klindt Myrvoll, Anders (1); Jackson, Andrew (2); Bingham, Nicola (2); Leikes-Rarugal, Carlos (2); O’Brien, Ben (3); Duncan, Sholto (3); Kreymer, Ilya (4); Ko, Lauren (5); Mulliken, Jasmine (6) |

• Still almost finished with the updated JWAT for validation of Warc-files - awaiting build for JAVA8

• Quite a few enquiries from researchers on our Facebook-content. We have a lot of old content, but curated new content is very sparse. There’s no good way to get Facebook content, cause our account will be recognized as a robot quickly, when using browsetrix cloud eg. and blocked or logged out. We are testing the limits with browser-profiles in Browsertrix cloud and logged-in crawling of Facebook - and it’s possible, but scoping will be important.

• NAS 7.4.3 in production

• SolrWayback updated 4 days ago - https://github.com/netarchivesuite/solrwayback/blob/master/CHANGES.md

BnF

The 2022 broad crawl has been launched on October 11th. According to our estimates, the harvest should reach a total size contained between 145 and 150 TB. It should last 5 and a half weeks and end between the middle and the end of November.

We plan to launch a final Video crawl this month. The departments of the BnF have been appealed to propose new selections. 243 selections were added between September 19th and October 27th and we are going to run our size estimates on 313 Youtube channels never collected at that time.

The last crawl of the IIPC War in Ukraine collaborative collection ended on November 3rd. Two others had been previously launched on September 22nd and July 26th respectively. A total of 964.4 GB of data have been collected for 1055 seeds selected by around thirty institutions.

ONB

BNE
On October 5th we held the workshop on non-print legal deposit and web archiving in the framework of the legal deposit working group of ABINIA (Association of Ibero-American States for the Development of National Libraries of Ibero-America). 13 countries and more than 40 people attended the workshop. It was interesting to exchange ideas on non-print legal deposit and web archiving and the problems they have to carry it out, mainly legislation, staff and resources. Countries like Chile, Peru and Colombia want to start archiving the web, all are considering NetarchiveSuit as a possibility.

The .gal domain has been harvested last month. A broad crawl of the regional domain of Galicia with a total of 6,525 domains and 285 GB.

KB-Sweden

Next meetings

- December 6th
- January 10th, 2023

Any other business?