2021-11-02 Statusmeeting

Agenda for the joint NetarchiveSuite tele-conference 2021-11-02, 13:00-14:00.

Participants

- BNF: Clara
- ONB: Andreas
- KB/DK - Copenhagen: Anders, Thomas, Stephen, Tue
- KB/DK - Aarhus: Colin
- BNE: José, Alicia, Miguel
- KB/Sweden: Peter, Jonas

Update on NAS latest tests and developments

Status of the production sites

Netarkivet

- Broad Crawl - finishing upload of step 2 harvest
- Event harvest starting up - Kommunalvalg 2021 (local elections)
    - Identifying candidates +8000!
- Upgrading to 7.3 NAS and Bitmagasinet (everything seems to run smoother compared to Bitarkivet). Hadoop-part almost ready.
- IIPC project proposal (awaiting IIPC still)
  User-Friendly High Fidelity Browser-Based Crawling for All - Proposal for IIPC Discretionary Funding Program 2021-2022 (1).pdf
- Twitter API-solution/pilot project - next steps close to be presented to management.
- Looking into NZ Web archives use of https://www.brandwatch.com/ to get Twiter, Instagram and Facebook-content (including comments) via API (as .XLS or CSV as I understood it)
- Setting up a Crawlogs and Gephi-system to find bottlenecks in harvests.
- Working on curator workflow/nomination pipeline with students from IT-University of Copenhagen
- Youtube-harvesting. Contact found via google Denmark and mails sent. Awaiting their feedback.
- Outsourcing harvest status
- Making a presentation for the Polish Webarchive initiative
- IIPC to use SolrWayback for collections
- Royal Danish Library part of IIPC Steering Committee

BnF

Our 2021 broad crawl was launched on the 11th of October. The chosen settings are 2100 URLs per domain, with a limit of 3 days per job. The crawl is due to finish in the middle of November and the budget should be around 112-115 TB.
At the start of the broad crawl, we had very slow jobs because of several million discovered URLs.
Some of our seeds redirect to a location like "http://fr/" or "http://com/". Heritrix considered "fr" and "com" as domains and added all the .fr or .com sites to the queue (a fix is ongoing on Heritrix: https://github.com/kris-sigur/heritrix3/commit/69b023199d3ad176b83c7e6d7dbb793c7a8adf66).

The BnF DataLab was opened on the 18th of October. It is a research assistance and support service set up by the BnF in partnership with the TGIR Huma-Num. The DataLab is intended for researchers who want to work on digital collections of the BnF.
A presentation about web archives was carried out by the digital legal deposit team on this occasion.
Moreover, a research project relating to web archives has been selected, among nearly 20 responses to a previous call for proposals launched by the BnF DataLab. This project led by Valerie Schafer is called "Buzz F, a history of online virality". The purpose is to reconstruct fleeting phenomena of online virality from traces found in the archives.

A new access to our "Archives de l'internet" will be opened at the Champs Libres Library in Rennes on November, 18th. It is the 21st access (out of 26) which will be opened in public libraries.

Finally, we will also organize a Webinar about web regional harvests, on the 9th of November. Up to now, three regional crawls are launched each year (Alsace, Lorraine and Languedoc-Roussillon). The aim is to exchange about these harvests and to develop new crawls with the other provinces.

ONB

BNE

In the next days a new regional broad crawl of the domain .gal (Galicia) will be launched.

We have participated in the international IIPC collection about Afghanistan regime

We have different technical problems with the new update that we have sent by email. In summary, the problems and questions are three:

- We don't know if in the new versión of NAS 7.2 is necessary to assign memmory to the process on each machine. Is It better if we do that?
- We have problems with setting of Postgres in WaybackIndexerApplication. Is there some documentation about this subject?
- In the result of OpenWayback we have strange results (3 crawls at the same time (same day, hour, minute and second)) and we don't know is it normal?

KB-Sweden

Next meetings

- December 14th
- January 11th, 2022

Any other business?

€€€€€€€