Update on NAS latest tests and developments

BNF is currently working on these features:

- NAS-2876 - Getting issue details... [STATUS]
- NAS-2877 - Getting issue details... [STATUS]
- NAS-2878 - Getting issue details... [STATUS]
- NAS-2879 - Getting issue details... [STATUS]
- NAS-2880 - Getting issue details... [STATUS]
- NAS-2881 - Getting issue details... [STATUS]
- NAS-2882 - Getting issue details... [STATUS]

Status of the production sites

Netarkivet

- Event harvest: shooting in Fields Shopping Mall started very early after the incident Sunday evening.
  - Using NAS/Heritrix, Twitter API and archiveweb.page
- Step 2 of the second broad crawl 2022 around half finished
- SolrWayback "live" QA still up and running and is great for QA.
- IIPC Browserbased-crawling project
  - We have an update meeting tonight and have had input during the IIPC GA-sessions.
  - Lots of user input from the Netarkivet team (curators, engineers and more to the Google Doc).
  - Great possibilities
  - Playback is important - browsers playing a bigger role with more advanced crawling/playback. As Kris put it: advanced crawling needs advanced playback
- Working on updated JWAT for validation of Warc-files ongoing
- Talks with the Norwegian web archive Nettarkivet - they use a browserbased crawler they made themselves called Veidemann: https://github.com/nlnwa. They are looking into SolrWayback for search/discovery (and maybe playback)
Last week, we launched our "Auction house" crawl, which concerns French auction houses websites. About 200 websites had been selected. Last year, we had been blacklisted by large auction sites. So we set up a specific harvest system for auction.fr where many websites are hosted. We added filters on all the other jobs in progress before starting the harvest and we created a special queue management to group the URLs of all hosts which belong to a website into one particular queue. This makes it possible to avoid sending too many requests at the same time as well as to limit the harvest to 100 000 URLs per website.

The LIFRANUM crawl carried out in partnership with researchers from the Jean Moulin University Lyon 3 and the Lumière University Lyon 2 is about to be launched.

The project aims to identify and map the corpus of digital French-speaking literature (sites, blogs, social networks). About 1100 sites will be crawled for this harvest with a specific budget of 15 000 URLs. The harvest should last about 1 or 2 weeks.

Finally, we are continuing the preparations for our 2022 broad crawl.

ONB

BNE

Catalonia has his own Project Padicat and his own system of harvesting. This month, for the first time, we are going to carry out the broad crawl of the .cat domain in colaboration with the Library of Catalonia.

Special harvesting for the LGBT pride in Spain, specially social networks

National Library of Spain continues to collaborate with the Barcelona Supercomputing Centre. They are going to make a second extration of data to create a new and improved version of MarIA, first massive artificial intelligence system in the Spanish language: https://www.bsc.es/news/bsc-news/first-massive-artificial-intelligence-system-the-spanish-language-maria-begins-summarize-and. We are studying different lines of action to apply AI to the Spanish web archive.

The National Library of Spain has a new website with new design, more intuitive and attractive: www.bne.es and our section: https://www.bne.es/es/colecciones/archivo-web-espanola

Different countries from Latin America have show interest in Web Archiving. This month, we’ll make two meeting with Peru and Chile. They want to know our way of working and tools, and Peru has shown interest in Netarchivesuite.

KB-Sweden

Next meetings

- September 6th
- October 4th
- November 8th
- December 6th
- January 10th, 2023

Any other business?