Update on NAS latest tests and developments

Status of the production sites

Netarkivet

- Broad crawl
  - 1st broadcrawl step 1 - 2023 still running - closing soon.
  - Issues with Hadoop, updating to RHEL8 and more
- Browserbased crawling for all IIPC-project still awaiting funding for development this year
- Pull requests for Browsertrix crawler behaviour and Instagram
- Anders writing on a blog post/update on Browsertrix-IIPC-project: Will use the closing of semi-upscale super market chain Irma as an example (crawling Facebook, Instagram, TikTok, Twitter and maybe some embedde videos)
- KB will focus more on Browsertrix-project the next month
- Focus on goals for 2023 and what we currently can't do with our 3.5-5.5 FTE working with web archiving - hoping to make a strong case for more resources
- Twitter API-harvesting stalled a bit - also awaiting new paid API-solution (9th of February)
- Browsertrix status on project and how KB have used/and might use it in the future by Anders from KB with BnF - 16th of February (Online meeting)
- Figuring out a way to visualize web crawling for KB's permanent photo exhibition (Gephi or maybe even browserbased crawling progress screen-recording)
- Data dumps - 3000+ PDFs, defacements from the Danish web (crawl times) and some lists, CDX-summary-like extraction of data for Janne /AU (Warncnet-project)
- SolrWayback 4.4.0 software bundle has been released
  - SolrWayback bundle release 4.4.0 can be downloaded here: https://github.com/netarchivesuite/solrwayback/releases/tag/4.4.0
  - https://github.com/netarchivesuite/solrwayback/blob/master/CHANGES.md

  "Visualization of search result by domain' can now be shown by day, week and month instead of only year. Same goes for the domain statistics in the toolbar. This is useful for recent collection that does not go back years. (see #270) Thanks to Leslie Bellony from BnF for implementing this"

BnF

First of all, this month, we are going to launch an internal project to improve several of our harvests. The project will run until July. It includes several parts:
- improvement of the harvest of social networks (Twitter, Facebook, Instagram)
- experiments with Browsertrix within the framework of our next internal harvesting workshop in March.
- improvement of the press sites harvest
- setting up Podcast and TikTok harvests.

At the end of January, Wayback version 8.10.0 has been released. This new version includes the publication of our new virtual guided tour concerning Artificial Intelligence.
This guided tour is made up of 13 themes. The topics covered range from scientific and technical applications of AI to ethical issues, and include the link between AI and art or human sciences.
The sites presented in the guided tour were selected for the Artificial Intelligence harvest launched for the first time in December 2020, but there are also older captures, some of them dating from the early 2000s.
On this occasion, a homepage of the "Archives de l'internet" on the subject of artificial intelligence has been republished.
A new Video crawl is running since January 26th. We are harvesting 13 Youtube channels for an estimated size of 4,8 TB.
The .eus domain has been harvested for the first time. A broad crawl of the regional domain of Vasque Countries with over 13,000 domains and 730 GB. It is a milestone for us because for the first time we have managed to save all the Spanish domains: .es, .gal, .cat and .eus.

We continue to have problems when we want to harvest Twitter in general. We have tested lowering the number of objects to 5,000, but when we want to save in the same harvest many accounts, only the picture of the first ones accounts are saved, the rest only the text. We have not found a solution to this problem.

Since mid-January, we have detected a new error harvesting Twitter. When we want to save a hashtag, trendy topic or search we get a 404 error, even though they exist on the web. We think that Twitter has changed some security policy.

Next meetings

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

Any other business?