2021-07-06 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 2021-05-06, 13:00-14:00.

Participants

- BNF: Auriane, Clara, Sara
- ONB: Andreas
- KB/DK - Copenhagen: Tue, Stephen, Anders
- KB/DK - Aarhus: Colin
- BNE: José, Alicia
- KB/Sweden: Pär, Peter

Update on NAS latest tests and developments

Questions: Will NAS 7.1 be released soon? When will we merge back to the official H3 release? [https://github.com/internetarchive/heritrix3/tags]

KB.DK: We are in the process of establishing our new production backend and expect to be ready to start installing NetarchiveSuite within a couple of weeks. We have made a number of minor improvements and bugfixes to our new backend functionality, but these are likely only of interest to ourselves. However this does mean that we will be making a 7.1 release very soon so please get your pull-requests in now!

7.1 will also feature some additional heritrix functionality which we pulled in directly from BL to our own Heritrix fork. Subsequently (and largely due to our queries about the subject) these additional features were also added to the community build of heritrix. But in the meantime our version had already diverged from the BL version. So we may not be able to make a smooth merge with the very latest community version in time for 7.1. But we can cherry pick anything essential such as the proposed revisit fix, or these can be made as pull requests to [https://github.com/netarchivesuite/heritrix3].

The new functionality we have grabbed from BL is:

1. sitemap extraction. Two extractors. One to grab sitemaps from robots.txt and one to parse them
2. json extraction: One extractor to find url's in Json files

Status of the production sites

Netarkivet

- We have begun to prepare our next broad crawl (the second in 2021) and we expect to start it september 2021
- We are delivering 2 data dumps in the coming months: one on pandemics from 2001-2021 and another about contact-information on websites!
- NAS 7.1 is tested and will be put into production after the summer vacation - mid-august
- Curators, developers and programme manager recommends the community-way vs. outsourcing harvesting. UKWA/Webrecorder makes proposal:
  - Towards a collaborative, open-source web archiving system - By Webrecorder, in collaboration with UKWA
  - IIPC Online Hours notes
  - [https://trello.com/c/kNXeWyrL/55-integrating-browsers-into-heritrix3]
  - [https://trello.com/c/yMgiUiP3/56-crawl-ui]
- Extractor YouTubeDL in production.
- Kb.nl wants to run NAS - should we invite them here?

BnF

Last week, we launched our "Auction house" crawl. The aim of this harvest is to crawl French auction houses websites and to archive information concerning current auction sales and auction results. This harvest is the result of a collaboration between the National Institute for Art History (INHA) and the BnF. About 200 websites have been selected.

Two weeks ago, we also launched two crawls, the first one is about Social movements and the theme of the second one concerns solidarity. There are respectively 977 and 476 selected websites.

During the summer, we will also realize a harvest about the Olympic and Paralympic Games, that will take place in Tokyo.

And finally, we began to prepare our next broad crawl. It will be our 17th broad crawl and it will be launched in October 2021.

ONB

BNE
• 1st July We have started our annual broad crawl of domain .es, we are using a new user agent: userAgentTemplate=Mozilla/5.0 (compatible; bne.es_bot; @OPERATOR_CONTACT_URL@) Firefox/57 the same we use for social networks
• Last week we reactivate an event crawls about “LGBT pride”
• Our Coronavirus collection reached 6000 seeds, more 90 TB of information about the pandemic

KB-Sweden

Next meetings

• September 7th
• October 5th
• November 2nd
• December 14th
• January 11th, 2022

Any other business?