2021-06-08 Status meeting

- Update on NAS latest tests and developments
- Status of the production sites
- Next meetings
- Any other business?

Agenda for the joint NetarchiveSuite tele-conference 2021-06-08, 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

BnF fix on revisits records missing: https://sbforge.org/jira/projects/NAS/issues/NAS-2870

+ tooltips on the monitoring dashboard.

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

BnF

Last month, we published a page about our Covid-19 collection on the BnF website. This web page, entitled "Covid-19 and the lockdown of March 2020 in the Web Archives", presents among others the harvest and the virtual guided tour published last March on our Web Archives. Moreover, a slide show with some captures will be added very soon.

You can consult this page at the following address: https://www.bnf.fr/fr/covid-19-et-confinement-de-mars-2020-dans-les-archives-du-web

Last May, the launch day of the ResPaDon project took place. The discussions were rich throughout the day. The speeches have been recorded and it will soon be possible to view them on https://webtv.univ-lille.fr/

For the second consecutive year, we are going to launch our two crawls about Environmental Issues and Artificial Intelligence. We plan to double the size this year: more than 650 websites for the first one and 450 for the second one have been selected.

We also began to prepare our Flash harvest that is scheduled for the end of June. We will crawl about a hundred websites with Flash animations through a semi-automatic process. This follows the harvesting workshop we organized last March.

ONB

BNE
• We have started indexing a part of our collections, and we are studying the resources we need to implement solrwayback in our entire collection, almost 1 Petabyte
• Since May we have problems to harvest Facebook profiles. Now when we want to harvest a profile, it redirects us to login page and we do not harvest any contents of this social network.
• We have incorporated two new librarians in our team
• This month we have two meetings with our regional web curators from different part of Spain. We will work on the quality control regarding web pages harvesting

Next meetings

• July 6th
• September 7th
• October 5th
• November 2nd
• December 14th
• January 11th, 2022

Any other business?