2019-12-03 Statusmeeting

- Update on NAS latest tests and developments
  - Cookie-bug
  - Heritrix Status
- Status of the production sites
- Next meeting
- Any other business?

Agenda for the joint NetarchiveSuite tele-conference 2019-12-03, 13:00-14:00.

Participants

- BNF: Clara, Sara, Géraldine
- ONB: -
- KB/DK - Copenhagen: Tue, Stephen, Anders, Kristian
- KB/DK - Aarhus: Colin, Sabine, Knud Åge
- BNE: Alicia, María, Manuel, José
- KB/Sweden: Par, Thomas, Peter

Join from PC, Mac, Linux, iOS or Android:

https://kbdk.zoom.us/j/104443571

Or an H.323/SIP room system:

H.323: 109.105.112.236
Meeting ID: 104 443 571
SIP: 104443571@109.105.112.236

Or Skype for Business (Lync):

https://kbdk.zoom.us/skype/104443571

Or Telephone:

Denmark: +45 89 88 37 88 or +45 32 71 31 57
United Kingdom: +44 203 051 2874 or +44 203 481 5237 or +44 203 966 3809 or +44 131 460 1196
Finland: +358 9 4245 1488 or +358 3 4109 2129
Sweden: +46 850 539 728 or +46 8 4468 2488
Norway: +47 7349 4877 or +47 2396 0588
US: +1 669 900 6833 or +1 646 558 8656
Meeting ID: 104 443 571
International numbers available: https://zoom.us/u/acRu0MV3xJ

You can join a meeting by using apps from a pc, a tablet or a smartphone, but you can also use the browser based version (it works with newer versions of Chrome or Firefox)

Please click OK if you see the system dialog.

Launching...

If nothing prompts from browser, download & run Zoom.
If you cannot download or run the application, go to https://zoom.us

Update on NAS latest tests and developments

Cookie-bug

BnF discovered that NAS 5.6 no longer sends cookies with any requests and this was having an effect on the quality of their harvest. The bug was easily reproducible and was found to be have been introduced between versions 5.4.2 and 5.5 so the Netarkivet production system is also affected.

After a considerable amount of detective work, we discovered that the bug came not from a change in NAS code or Heritrix code but from a change in one of the 3rd Party libraries - specifically we had somehow come to downgrade the version of guava shipped with NAS. Simply substituting a more recent guava version in the bundler zip makes the issue go away. What we have not done:
1. We don't know exactly why the build started packaging only the older guava version
2. We don't know what is wrong with the older guava version that it causes this behaviour, and most importantly
3. We haven't released a patch. We should really release patches to both 5.5 and 5.6 branches.

Heritrix Status

Colin and Clara have spent some time analysing the NAS modifications to Heritrix to see what would be needed to get a community-version of Heritrix that we could use in NAS. There seem to be three changes we would want in:

1. Adding a timeout to the crawler-trap regex test
2. Adding a filter to prevent inline images being interpreted as links
3. Modifying the frontier to add additional methods to browse the queued URLs (including upgrading Berkeley DB - for which Andy Jackson already has a pull request).

Colin has reimplemented each of these in separate branches on top of the current IIPC/IA master branch (including Andy's pull request for nr. 3), and created a fourth branch (https://github.com/netarchivesuite/heritrix3/tree/h3.4-merge) in which all three modifications are merged. There is also a Netarchivesuite branch (https://github.com/netarchivesuite/netarchivesuite/tree/h3.4) which can be built against this heritrix once the heritrix has been installed locally with maven. What we need to do now is:

1. Basic functional testing of nas/h3.4 (ie the release candidate for NAS 5.7)
2. More extensive (acceptance) testing of nas/h3.4
3. Make a series of pull requests to try to get our code into the main Heritrix repository.

Even if we aren't able to get our pull-requests accepted quickly, we should still base future releases on this work as we would then have a Heritrix version very close to the community version, making it easy to pull in future upstream changes.

Status of the production sites

Netarkivet

Our 4th broad crawl for 2019 is running (included the selective crawls: very big sites, ministries and administrative bodies, scientific publications, YouTube)

We focus on extremists sites, we are working on the identification of left wing, right wing and Islamic extremist sites.

We are discussing how to give researchers access to our documentation. When we chose Confluence for our documentation we thought of giving researchers access to relevant pages, but this does not seem to be a good idea. netarkivet.dk will be the platform for giving access to relevant documentation.

Other projects keeping us busy (almost the same as last month):

- Work on risk assessment
- Implementation of SolR Wayback
- Consolidation and upgrade of BCWeb (build up a community)
- Revision of collection strategies
- Capture of content behind paywalls – the never ending story
- New procedure for access (both internal and external)

BnF

ONB

BNE

Currently, these are our main tasks:

- We are working to install BCWeb 6.1. We think it will be ready next week.
- The IT Team is working to resolve the index problems that we have in OpenWayback
- We are considering doing a “massive” crawl of periodicals in free access.
- We create a new collection on climate change because of the COP that is taking place in Madrid.

KB-Sweden

Next meetings
Any other business?