2020-02-04 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 2020-02-04, 13:00-14:00.

Participants

- BNF: Clara, Sara, Géraldine
- ONB: Andreas
- KB/DK - Copenhagen: Tue, Stephen, Anders, Kristian
- KB/DK - Aarhus: Colin, Sabine
- BNE: Alicia
- 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 8833 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...

Update on NAS latest tests and developments

In Denmark we have two new developers working, for the moment, full-time on webarchiving - Rasmus and Peter. The major effort is on replacing our current backend which will require

- Reimplementing access (getRecord) to in a secure way that avoids the need to go through JMS and ftp
- Reimplementing all essential batch jobs in more modern mass-processing framework ie hadoop

As far as Heritrix 3 work is concerned, the main issue is what to do with our various homebrewed heritrix extensions. For each extension we should either

- Retain our extension (the default)
- Move to the newest equivalent Heritrix extension
- Merge our extension with the latest heritrix version
In each case we will need to analyse the code to decide what makes sense.

**Status of the production sites**

**Netarkivet**

**Broad crawl**
We postpone our first broad crawl for 2020 for having time to “clean up” after the last broad crawl for 2019 and for having time for a rigorous preparation of the next crawl.

**Event crawl?**
We almost were getting a new “Cartoon Crisis”, as a cartoonist from Jyllandsposten designed a Chinese flag with Corona virus instead of yellow stars. We kept an eye on what would happen and finally we decided, that we could do with a special crawl on foreign media reactions, just to supply our regular selective news media crawls.

**Access**
We have big trouble with the performance of our replay tools
- Viewerproxy: no https support
- Citrix Wayback: much to slow
- Open Wayback: an upgrade to IIPC’s new 2.4 release might help
- Blacklight (full text search): not all of the facets are applicable

Hopefully we soon will get solved legal issues so we can implement SolR Wayback.

We are finalizing a new application form for access to Netarchive. At the same time we work on a new modernized procedure for access, not least in consideration of GDPR

**BnF**

**ONB**

**BNE**

**Event crawl**
Finally, we have a new Government, so we finished our event crawl of Spanish Government elections.

We are planning to create a new event crawl of videogames this year

**Broad crawl**
We are going to launch this month a “massive” crawl of periodicals in free access. We want to harvest more than 10,000 websites that hosted electronic serials.

**NAS and BCWeb**
We downloaded BCWeb version 6.1, we made all the changes that we did in our last version, and now, we want to install it in a test environment to see if everything works ok.

We solved part of the problem that we had with de indexation of NAS. We can see the captures since January 2020, although we still can’t access to the captures of 2019

**10 years of Spanish Web Archive**
February 20, it will take place a conference to celebrate 10 years of Spanish Web Archive. We will review the state of the art of different web archives in the world, the challenge of its long-term preservation, as well as the use that researchers can give to this set of information.

**KB-Sweden**

**Next meetings**
- March 3, 2020
Any other business?