2020-09-08 Statusmeeting

Agenda for the joint NetarchiveSuite tele-conference 2020-09-08, 13:00-14:00.

Participants

- BNF: Clara, Sara, Alexandre
- ONB: Andreas
- KB/DK - Copenhagen: Tue, Stephen, Anders
- KB/DK - Aarhus: Sabine, Kristian, Colin
- BNE: Alicia
- KB/Sweden: Pär, 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):

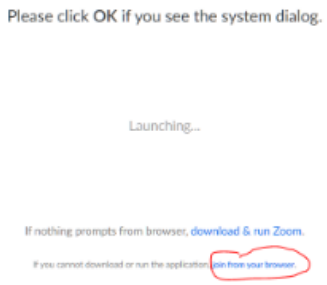
https://kbdk.zoom.us/skype/104443571

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)



Update on NAS latest tests and developments

Any feedback on NAS 6.0 ?

Status of the production sites

Netarkivet

Broad crawl

We started our second broad crawl for 2020 on 20 August, the first step with a byte limit of 50 MB finished on 2 September. On 21 August we started the separate crawl of ultra big sites, this crawl is still running.

Event crawl

We have to decide, whether we want to stop the event crawl on Corona in Denmark or not, there are different opinions on that issue.

Miscellaneous

Everything is prepared for the French trainee: we signed a contract and he will start on 28 September. He wants to work on visualization of data and Netarchive.

We started a collaboration with the IT-University in Copenhagen: students participating in a course on project work and communication for software developers will work together with us on several special challenges.

We try to solve various technical issues; we got aware of most of them on the base of emails from persons dealing with certain web sites. These issues are for example:

- URL's which do not change, when you click on links from the front page (gaffa.dk)
- Embedded tables (dfi.dk)
- Sites where we need a JavaScript interpreter for to render the pages (rehpa.dk)

We are going to look at the new features in BCWeb on an installation in a test environment

BnF

After the upgrade of NAS and Heritrix in June, we have observed the evolution of the QA indicators by comparing similar jobs run before and after the upgrade. The findings are positive : for a same job type, we crawl more URLs with less 404 errors with the new version, and the improvement is particularly significant with the image files, with a growth of the number of crawled images between 19 % and 98 % depending on the different types of jobs. We are very happy with this quality improvement, however we have to manage with larger WARC files and to reassess our budget estimate. Our annual broad crawl will be launched in October and we have to carefully adjust the parameters in order to comply with budget forecast.

The new version of BC web (7.3.0), with new functionalities such as duplication of records and improvement of the advanced search and of the deduplication, has been successfully put in production at the end of July.

ONB

BNE

Broad crawl:

This year the result of the broad crawl has been 1.930.000 web sites (around 50 terabytes of information). The number of domines has increased but the information published on internet has been less than the year before. Of all web sites that have been saved a 87 per cent have been fully and completely recollected.

Covid19 Collection:

On the other hand we continue with the collection about Coronavirus which increases each week. Actually it contains more than 4000 (four thousand) web sites

KB-Sweden

Questions:

Do you treat certain types of web sites/domains as uninteresting to harvest, and limit their budget or reduce the harvest in other ways? If yes:

- Which categories of web sites?
- How do you identify the category and find which web sites to treat specially?
- How do you reduce the harvest there – data limit, object count limit, reject rules?

We would like to avoid the very large amount of web sites containing huge product catalogues, often with lots of images on each product. But are there ways to do find and avoid/limit them in some (semi-)automatic way?

(On the wish list – when you have identified such a site – would also be a way to harvest a specified proportion of it, e.g. 1 %, randomly selected among a representative selection of different types of pages … J)

A side-track to this is more complicated crawler traps which often show up on these (and other) sites, e.g. infinite loops of types which Heritrix can't detect (a/b/c/a/b/c, pages referring to themselves with extra parameters etc.). Hints?

Next meetings

- October 6, 2020
- November 3, 2020
- December 8, 2020
- January 5, 2021

Any other business?