2017-02-07 Statusmeeting

Agenda for the joint BNF, ONB, SB, KB and BNE NetarchiveSuite tele-conference 2017-02-07, 13:00-14:00.

Practical information

- Go to https://c.deic.dk/netarkivstyregruppe **or do we use Zoom now**?
- Login as guest
- Write your name
- Insert password: wayback

Participants

- BNF: Lam, Géraldine, Sara
- ONB: Michaela, Andreas
- KB/DK - Copenhagen: Stephen, Tue, Birgit
- KB/DK - Aarhus: Sabine, Colin
- BNE: Mar
- KB/Sweden: -

Upcoming NAS 5.3 Release

Status of developments.

On BnF side, 2 pull requests have already been submitted (see details)

PR1: https://github.com/netarchivesuite/netarchivesuite/pull/36

PR2: https://github.com/netarchivesuite/netarchivesuite/pull/37

(These have been validated by our automatic system test. CSR)

PR3: coming soon. It will also include the following fixes:

H3 pauseAtStart bean property ignored by Harvest Controller NAS-2596 - Getting issue details... STATUS

WARC-Refers-To-Date in WARC revisits records do not have the right original record date:
NAS-2602 - Getting issue details... STATUS

(Release Date? Next week (week 7) is a school holiday in DK. Colin would like to start work on organising release 5.3 immediately after that, if PR3 is available. So early March - week 10 - for release.)

NAS workshop in Vienna

Date and participants: http://doodle.com/poll/mvgm5w2v3bk6dsc7

Review of possible topics: 2017 NAS workshop

Questions from KB

1) What CDX format are you using today and plan to support within next year?

2) Which version of (Open)Wayback are you using today and what do think about the future development of OpenWyback?

3) Which social media can you archive today?

Status of the production sites

Netarkivet

We are concentrating our efforts on the capture of social media. Last week all curators met physically to kick off the discussion of our social media strategy. We started with analyzing the social media in order to be able to decide the selection and to propose a crawl frequency.

Some statistics for 2016: There are 1.097.585 active websites listed in NAS. 180.046 sites are bigger than 10 Mbytes in 2016. We harvested ca. 27 milliard objects. The total of the archive in the end of 2016 is 769 TB; we harvested 95 TB in 2016, that is to say 35 TB less than in 2015.

Development is focused on migrating the existing archive to compressed (.gz) format. Compressing the files is easy - the difficulty is finding and updating all references to the old files:

- in metadata files
- in cdx indexes
- in the admin database
- in the checksum database

... on a running system, and with minimal downtime.

BnF

We are working what should be the final corrections before changing over to Heritrix 3 and NAS 5. Our tests have not identified any major problems, we are continuing to analyse the results to prepare for any changes that might arise, such as an increase in the amount of data collected. Once we start crawling with H3 we will increase our usual monitoring to be able to deal with any unexpected changes.

We started our 2017 elections crawls.

ONB

- At the end of January we could finally finish our presidental elections crawl.
- We are currently preparing our 5th domain crawl
- Finally we redeployed our Testenvironment. For that, we made some convenient changes in the DeplyApplication (deploying with optional logo images). See our Pullrequest https://github.com/netarchivesuite/netarchivesuite/pull/38

Answers to Questions from KB
1) We are using the CDX-Format coming from WaybackCDXExtractionARCBatchJob
2) We are using currently OpenWayback 2.3.1
3) We crawled some facebook pages by using https://webrecorder.io

BNE

We've just opened a new web collection for all the regional web curators in Spain to participate. It is a daily collection about newspapers. We already have a daily newspapers collection, but only for national media. This new one is for regional newspapers.

The regional web curators are getting more and more involved in the management of their own collections, adding sedes in CWeb and doing Quality Assurance.

We are planning our yearly domain crawl (maybe around april), but its launching is related to the implementation of NAS 5, which is up to our engineers.

Our main and closest goal is to give access to users to our web archive, not only at the Library computers, but also at the regional libraries. Once we are sure that the security measures are implemented regarding the legal constraints at every access point, we'll open it and let you know. We are looking forward to that momento and also a little afraid of it. This is expected to happen in 3 or 4 weeks as máximum.

The non-print legal deposit team is expected to be reinforced with more people in a couple of weeks.

KB/Sweden

Next meeting

March 7th

Any other business?