2016-05-10 Statusmeeting

Agenda for the joint BNF, ONB, SB, KB and BNE NetarchiveSuite tele-conference 10-05-2016, 13:00-14:00.

Practical information

- Go to https://c.deic.dk/netarkivstyregruppe
- Login as guest
- Write your name
- Insert password: wayback

Participants

- BNF: Sara, Annick
- ONB: Michaela, Andreas
- KB: Søren, Tue, Jonas, Stephen, Nicholas
- SB: Colin, Sabine
- BNE: Mar

NAS 5.1 Update (Tue)

Now in use in Netarkivet production environment

Stephen made some small reports on what our template changes are, explaining issues and how our NAS4 broad/selective crawl are arranged.

KB/SB are planning on changing our strategy on how we work with broad crawls now and will change focus towards more selective harvest.

This hasn't been agreed, so this is not described anywhere yet.

Regarding our new H3 templates we have from start been using H3's own default template .

This template has then been modified to work as closely to our old H1 templates.

To compensate for some issues we had in H1, new changes has been made and described in the report 'Changes in H3.docx'.

Generating jobs in NAS5_1.docx, Issues in H3.docx

IIPC GA (all)

Feedback and important information from GA

NAS workshop (Sara)

Topics

1) Share experience with NAS 5 and Heritrix 3

2) Discuss challenges with specific types of sites (news, social media)

3) Discuss collection strategies

4) Discuss features/a GUI to handle the harvester

5) Look into the possibility to integrate another crawler into NAS (Colin proposed to come with a prototype with a headless browser)

Schedule

End of January 2017 - 2,5 days - in Vienna

Poll from Michaela http://doodle.com/poll/nk6dfc3kav4a4hs8

Status of the production sites

Netarkivet

- We have moved our production site to **NAS 5.1 H3**
- We will start the second **broad crawl** 2016 as soon as NAS 5.1 and Heritrix 3 are running “smoothly”
- The **event crawl** on the **refugee crisis** is stil ongoing**:** As it is a supplement to our selective news media and social media crawls, it is a very little event crawl.
- We will participate in the IIPC collections **Olympics 2016** and **Online News around the World: A snapshot in Time**
- We are preparing for a new event crawl on the European Capital of Culture project “Aarhus 2017”: we are looking at different scenarios for this event crawl
- We are still unable to harvest anything from **Facebook**.
- We are revising our **collection strategy**: There will be less broad crawls and more selective crawls. At the moment we are looking at the selective news media crawls. According to our resources we need a more streamlined approach for an extended number of domains to be crawled
- The social platform **arto.com** will be closed down at june 1st . We were offered a private crawl of the entire site (no WARC files, but likely WARC compatible). We decided to say no thanks and to do a last crawl of the entire site on our own.
- We are working on a business model (juridical and financial issues) for giving corpora from Netarchive to research institutions. Our first customer will be the University of Southern Denmark.

BnF

- We are still running our ongoing selective crawls (the biggest one is annual focused on big hosts and domains, social movements).
- We installed Java 1.7.79 on some harvesters within a specific channel to solve HTTPS problems for specific crawls (news and official publications).
- We are still working on our Corpus project.

ONB

- Please see doodle poll http://doodle.com/poll/nk6dfc3kav4a4hs8 to select the dates for the NAS meeting in Vienna (25.-27.1. or 30.1.-1.2.)
- Event crawl about presidential elections is still ongoing. One of the political parties is blocking our crawlers. We used webcrawler.io to archive the website. We still need to find a way to integrate the content into our webarchive.
- We completed an important milestone: the new user interface for the inhouse webarchive terminals has been launched. It includes a partial fulltext search, screenshot preview and uses Open Wayback. The next step will be going online, archived content will not be accessible, search-function only.

BNE

- The first .es domain crawl is running since April 4th. Our engineers estimate it will last until the end of July or mid August.
- We are trying to connect BCWeb to NAS development environment to give access to the web curators from our regional libraries.
- As our General Elections are going to be repeated in June 26th, we didn't close yet the General Elections event crawl that started on December 2015. Web curators from the regional libraries are nominating seeds for this collection.
- At the moment, the web archiving team is even smaller than it was. If the librarians were two (Sole and me), I'm now on my own, because Sole moved to another position at the Library. We are trying to recruit more people for the team, but so far the situation is even worse than it used to be.

Next meeting

2016-06-14

Any other business?