2012-01-10 Statusmeeting

Agenda for the joint BNF, ONB, SB and KB NetarchiveSuite tele-conference December the 10th 2012, 13:00-14:00.

Practical information

- Skype-conference
    - Mikis will establish the skype-conference at 13:00 (Please do not connect yourself):
        - BNF skype name: sara.aubry.bnf
        - ONB skype name:
        - SB skype name: mss.statsbiblioteket.dk
        - KB skype name: christen.hedegaard.kb
    - TDC tele-conference (If it fails to establish a skype tele-conference):
        - Dial in number (+45) 70 26 50 45
        - Dial in code 9064479#
- BridgIT:
    - BridgeIT conference will be available about 5 min. before start of meeting. The Bridgit url is konf01.statsbiblioteket.dk. The Bridgit password is sbview.

Participants

- BNF: Sara
- ONB: Michaela and Andreas
- KB: Tue, Søren and Jonas
- SB: Colin and Mikis

- Any other issues to be discussed on today's tele-conference?

Followup to workshop

- Actions.

> - Peter, Michaela and Sabine will send individual posts to the curator mailing list with news on the status at the respective sites. Tue will talk to Sabine about compiling the update.
> - Mikis will try to get an overview of the added curator wiki content and send a mail to Sara, who will create a post to the curator mailing list regarding the new content..
> - Mikis and Sara will look into defined general Jira usage. Karen will contacted afterwards for analysis of NASC-17@jira.
> - The wiki content regarding templates and crawler traps isn't finished yet, even though some content has been added.

Iteration 50 (3.19 Development release) (Mikis)

- Codefreeze 15.feb.
- 3.19.0 release test

Status of the production sites

- Netarchive (Tue):

> Our 3th broad crawl in 2011 finished in December the 19th. 28 TB ran through our harvesters and only 15 TB was uploadet after 45 % deduplication.
> Max upload peek during the harvest has been 1,3 TB/day and we have tried in a stresstest to upload 3,3TB/day. So we have plenty of upload capacity.
> We are moving to NAS 3.18 this month and we plan to move to postgresql within the next 4 month.

- Netarchive - curator update (Sabine):

Broad crawls.We finished our third broad crawl for 2011 at 2011-12-19, it lasted 51 days, we harvested 27,9 TB/ 647.588.162 objects.

Our second broad crawl lasted 59 days, we harvested 27,5 TB/630.022.496 objects.

We decided to wait with the first broad crawl 2012 until February and in the meantime we are doing a special harvest on the biggest Danish sites and on the Danish ministries and administrative offices.

In our broad crawls 2011 we harvested a total of 79,3 TB og 1,8 billion documents/url's. In comparison to that the results for our selective crawls are peanuts ;o)

Selective crawls: We are working on improvement of our documentation: a draft of an overall collection policy for our selective harvests is available for presentation to our editorial board.

Event Harvests: From the Netarchive point of view the end of 2011 wasn't rather eventful.

Anyway – we did an effort on intermediality: Every year in week 46 – which is called a "usual media week" (Nov. 14-20 in 2011) SB collects programs from Danish local and amateur tv and radio stations. This year we decided to make an event harvest at Netarchive of those tv-stations, who are streaming their programs. Of cause, we could not capture the streaming, but as we harvested their sites daily, their program descriptions can supply our tv/radio collection.

We just started an event harvest, which will last for half a year: a harvest on the Danish EU presidency 2012

A wayback machine as you know it from archive.org is ready for our users.

- BNF (Sara):

We finished our 2011 broad crawl on December 26th. Main figures are: 1 billion of harvested URLs, 32,6 TB of data, duration of 11 weeks. We are focusing developpment efforts on BCWeb in order to start focused crawls on French presidential and general elections.

- BNF - Curator update (Peter)

Harvesting with NetarchiveSuite:The year 2011 was the first one when BnF used only NetarchiveSuite to organize all its crawls. We managed one broad crawl (as in 2010) but we also started a daily crawl for newspapers and focused crawls with differents schedules. We finished the year with 1.6 billion of harvested URLs (1.2 billion in 2010) for a compressed weight of 57 To (43 To in 2010).

The main challenge for 2012 will be to manage crawls for the French presidential and general elections. We will notably try to collect Twitter several times a day.

Snapshot harvest at the Bibliothèque nationale de France: As announced in October 2011, we finished our snapshot harvest on December 26th. These are the three main figures : 1 billion of harvested URLs, compressed weight of 32.6 To, duration of 11 weeks. The crawl went well except that we had an incident between step 1 and step 2 : we had less domains registered for step 2 compared with 2010, although the number of seeds was bigger. This was due to a parameter which calculates the number of generic errors of URLs (errorpenalty in the order.xml) : we commonly used it when we only had Heritrix but this time, with NetarchiveSuite, the report indicated only 999 harvested URLs instead of 1,000 when one error occurred. So we had two steps 2 : the first for 49,000 domains (for domains that had reached the limit of 1000 URLs in NAS), the second for 53,000 domains (for domains that had failed to reach the limit of 1000 URLs, but instead had 997, 998 URLs and so on due to the errorpenalty value).

During the monitoring, we noticed a large number of sites on e-business, online directories and ads, municipalities websites. We decided to stop collecting websites which last too long (about 3 to 5 per job): for step 1, this meant generally after 1 day; for step 2, generally after 4 days.

BnF would be interested to have some information about your own observations on your 2011 broad crawl.

New curator tool : BnF Collecte du Web (BCWeb):Since April 2011 we have been developing a curator tool for use by our network of 80 subject librarians at the BnF, with the aim of opening it up to external partners, notably for the election crawls. The tool, known as BCWeb, organises URLs using the types of collection we already use, by thematic departments and projects. It allows selectors to define the parameters to be used by NAS (depth, frequency and budget) as well as documentary fields such as keywords and notes. Search and browse features allow users to keep their selections up to date. An administration module allows the web archiving team to load the data into NAS: adding new URLs to Harvest Definitions and updating or deleting old ones.

BCWeb has been developed using the "Scrum" method of agile software development. This meant we were able to develop the main functions early on, but in the past couple of months we have encountered problems with performance and the integration of graphic elements of some pages. We have loaded the complete database of almost 14,000 URLs and are finalising the full-scale tests of the import into NAS. We aim to have the tool online by February, to allow the selection of sites for the presidential elections in April.

- ONB (Michaela):

New collection "Austrian Literature" was started in December, will be crawled monthly. Selection was done by Literature Archive of ONB.

Domain crawl is still running (stage 2), we will move to NAS 3.18 when crawl has finished

In December we gave access to the 2nd external library, the Federal Administrative Library

Date for next joint tele-conference.

- February 14th 13-14.

Michaela will not participate

Any other business