2016-07-12 Statusmeeting

Agenda for the joint BNF, ONB, SB, KB and BNE NetarchiveSuite tele-conference 12-07-2016, 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, Lam
- ONB: Michaela, Andreas
- KB/DK: Søren, Stephen, Nicholas
- SB: Sabine
- BNE: Mar, Juan Carlos, Fernando, Elena

NAS workshop in Vienna

January 30th 2017 - February 1st 2017 - Vienna

Please complete Michaela's poll : http://doodle.com/poll/nk6dfc3kav4a4hs8

IIPC crawler hackathon in London

September 22-23.Topics: 1) archiving with Warcproxy, 2) browser-based crawling, and 3) Web archiving APIs.

Is anyone attending?

- Olga from IIPC has invited BNF to participate. If anyone else is invited/will participate, please inform the NAS community to Exchange ideas before the hackaton.

NAS 5.2 Developpement Update

Feedback from KB/SB.

- ongoing Work, coordination with Kristinn form Iceland. All changes will be in the NetarchiveSuite code base. Søren sends his coding to Sara for reviewing. Coding is suposed to be finished in the end of July, august will be used for tesating, release in the end of August.

BnF getting started to migrate to NAS 5 and H3. Need help to get started.

- Everybody, who needs help: please use the developpers mailing list (not individual emails), so all can help each other.
- The key is using the right templates

Status of the production sites

Netarkivet

- The second broad crawl 2016 (with the limit of 100 MB per domain) finished at June 28. We harvested 11.255.368.320.635 bytes / 242.114.319 objects. We had problems with upload capacities at SB. We have worked out an action plan which will be implemented soon.
- We started an event collection for the Olympics in Rio 2016 on July 24. We also participate in the IIPC Olympics collection
- We are going to use our Archive-IT account to try to capture Facebook profiles.
- As part of our new collection strategy we have started working with university repositories, educational and law portals:
    - Research databases: We started with the collection of the Danish "PURE-repositories" including local hosted publications (as for example from JSTOR or Elsevier).We use our OAI-PMH-harvest-definition, which still is under optimization.
    - Educational portals. We are establishing contacts to the providers for to make agreements for harvesting login content.
    - Schultz Law portals: we have got login information from the publisher Schultz and after summer holidays we will assess the best method for collection.
- Our dissemination policy and strategy are getting the last brush up.
- A revised SB and KB's collaboration agreement on Netarchive has been signed of the directors from both institutions.
- We have finalized a recommendation on the compression of the WARC files in Netarchive.
- NAS 5.2 will be released soon.

BnF

Each year, the different sections of the BnF legal deposit department give a view of the documents they have received. L'*Observatoire du dépôt légal : reflet de l'édition contemporaine* is now available online (in French only): http://www.bnf.fr/fr/professionnels/depot_legal_definition/s.depot_legal_observatoire.html.
It gives analysis and raw data from 2015 on seed domains (more than 900,000 have appeared since the previous year and more than 500,000 have disappeared), on format, on http response codes, on the biggest harvested domains…

This month we also have several project crawls on different themes.

Among these project crawls, the annual one dedicated to French Official Publications is still going on with few new aims. Launched in the middle of June, it contains a sample of the web social presence of the central administration, with the decision to add the social media accounts of ministers and public bodies. While this is unfortunately without crawls of Facebook pages because of the now well-known problem of captchas, the goal is to reflect this type of official communication that was previously not so well covered in our selections. The frequence of the crawls of these specific ways to promote official publications, administrative and political communication could be extended in the future. The traditional aim of collecting the "classic" online publications is still relevant, with more than 800 URL seeds of traditional websites, crawled with a 100,000 URL budget for each.
Our annual crawl of auction houses has just finished. The scope of the collection is the same as in previous years, but last year the platform auction.fr, which represents about a third of the crawl, blocked access by our robots. The librarian in charge of the selection contacted the site owner who was happy to let us crawl the site, and the quality seems much better this year. We also have to be careful as the majority of the sites are hosted on two platforms (auction.fr and Drouot), and their catalogues and images are stored on a small number of hosts - we have to increase the budget for these hosts to collect as much as possible.

We are also maintaining our crawl "Solidarities" with the same scope as last year, though we have also included sites that were selected for an emergency crawl on the refugee crisis .

ONB

- Michaela has changed position within the library and since July 1st is head of the Digital Library. Her post will not be replaced at the webarchive. At the moment ONB is not sure what the NAS contribution will look like in the future. We will work on a new concept and allocation of tasks. Michaela (and of course Andreas) will still be the contacts for webarchiving.
- Please complete Doodle poll for Vienna meeting until end of July http://doodle.com/poll/nk6dfc3kav4a4hs8
- Crawl about presidential elections is still ongoing, the repetition of the election will take place in October.

BNE

- The first .es domain crawl, run with NAS at the Library finished on July 6th. It started on April 4th, so it took 3 months. From a list of 1.800.000 registered domains, only around 800.000 are active. The result is around 20 Tb and 460.000.000 objects. We fixed a limit of 100 Mb per seed and around 87% of the domains have been crawled entirely.
- The General Elections took place for a second time in June as the Parliament coming from the December 2015 elections didn't manage to designate a Prime Minister. So our General Elections event crawl launched by the beginning of December 2015 hasn't finished yet. So far, we have collected around 10,5 Tb. The regional web curators that collaborate in the project have been nominating seeds for this event crawl.
- The regional web curators are testing BCWeb on a preproduction environment and they are starting to manage their own web collections using this application. So the production environment of BCWeb is only managed by the National Library team so far. We hope they get the training and knowledge enough to start using the production environment by next autumn. In the meantime National Library web archiving team is launching some regional web collections of limited scope.
- A couple of weeks ago we welcomed two fellows at the team. They will be working with us for one year. Miriam is an information and documentation specialist and Elena is engineer.

KB/Sweden

Next meetings

- August 23
- September 20
- October 25
- November 29
- January 3, 2017

Any other business?