2015 29-30 January NAS workshop

Location: National Library of Estonia, Eesti Rahvusraamatukogu / Tõnismägi 2, Tallinn (http://www.nlib.ee), meeting point: main entrance

Participants:

Organization	Technical	Curator
BnF	Iam mai, Sara Aubry	Clément Oury, Annick Le Follic, Géraldine Camile
ONB	Andreas Predikaka (participating through Skype)	Michaela Mayr (participating through Skype)
DK	Mikis Seth Sørensen, Søren Vejrup Carlsen, Tue Hejlskov Larsen, Per Møldrup-Dalum	Ditte Laursen , Sabine Schostag, Ulrich Karstoft Have
Estonia	Meelis Mihhailov, Rando Rostok	Jaanus Kõuts, Tiiu Daniel, Elis Karpov Liina Abner (ebooks/newspapers discussion)
Spain	Juan Carlos García Arratia	Mar Pérez Morillo

- Topics to be discussed:
- Agenda
  - Schedule for Day 1 (Thursday 29)
    - 09:00 - 09:30 Welcome and coffee
    - 09:30 - 10:00 Workshop introduction (Sara Aubry)
    - 10:00 - 11:15 Institution updates and plans for 2015
    - 11:15 - 11:30 Coffee break
    - 11:30 - 13:00 Statistics on web archives using ISO metrics (Annick Le Follic)
    - 13:00 - 14:00 Lunch
    - 14:00 - 14:10 Introducing Heritrix 3 in NetarchiveSuite: NAS 5.0 status and plans (Mikis Seth Sørensen)
    - 14:10 - 14:30 Quick demonstration of Heritrix 3 (Søren Vejrup Carlsen)
    - 14:30 - 14:45 Introducing Heritrix 3 in practices: BnF approach (Sara Aubry)
    - 14:45 - 17:00 WARC track: WARC usage in NAS compared to Archive-it (Tue Hejlskov Larsen)
    - Coder track: NAS 5.0 code redesign (Mikis Seth Sørensen), Completed and remaining tasks (Søren Vejrup Carlsen)
    - 14:30 - 17:00 Curator track: monitoring and QA crawls with Heritrix 3 (Annick Le Follic and Géraldine Camile)
    - 15:30 - 15:45 Coffee break
    - 17:00 - 17:30 Tour of web archiving activities in Estonia (Jaanus Kõuts)
    - 19:00 - Dinner
  - Schedule for Day 2 (Friday 30)
    - 09:00 - 09:30 Harvesting complex websites: experiments with Archive-it 4.9/5.0 using 3.3.0 with Umbra (Tue Hejlskov Larsen)
    - 09:30 - 11:15 Digging in the data mines of the Net Archive (Per Møldrup-Dalum)
    - 11:15 - 11:30 Coffee break
    - 11:30 - 13:00 Heritrix tracks sum-up, review of NAS curator roadmap, community next steps (Sara Aubry)
    - 13:00 - 14:00 Lunch
    - 14:00 - 15:30 Ebooks/newspapers: deposit or FTP harvesting
    - 15:30 - 15:45 Coffee break
    - 15:45 - 17:00 Open space for an additional topic, individual discussions, workshop closing

The day before the workshop itself 6 NAS participants will give talks on the International Seminar on Web Archiving in Tallinn, see 2015-01-28 International seminar on web archiving in Estonia for details.

Topics to be discussed:

Heritrix 3 - technical	Heritrix 3 - curatorial	NetarchiveSuite	
• State of the art of developments, scope of 1st release, challenges • Upcoming developments: what, who, when • WARC format in Archive-it compared to NAS	• Feedback on testing • Missing features • Priorities for future developments	• Broad crawls: improve quality, reduce storage • Statistics based on ISO metrics • Ebooks/newspapers: deposit, FTP harvesting	

Agenda

Schedule for Day 1 (Thursday 29)

Location: *Cupola Hall*

09:00 - 09:30 Welcome and coffee

09:30 - 10:00 Workshop introduction (Sara Aubry)

10:00 - 11:15 Institution updates and plans for 2015

Update from ONB (Michaela Mayr)

Update from BNE (Mar Pérez Morillo)

Update from Estonia (Jaanus Kõuts)

Update from BnF (Clément Oury)

Update from SB/KB (Ditte Laursen and Sabine Schostag)

11:15 - 11:30 Coffee break

11:30 - 13:00 Statistics on web archives using ISO metrics (Annick Le Follic)

NAS_qual outcome sample in English

13:00 - 14:00 Lunch

14:00 - 14:10 Introducing Heritrix 3 in NetarchiveSuite: NAS 5.0 status and plans (Mikis Seth Sørensen)

14:10 - 14:30 Quick demonstration of Heritrix 3 (Søren Vejrup Carlsen)

14:30 - 14:45 Introducing Heritrix 3 in practices: BnF approach (Sara Aubry)

14:45 - 17:00 WARC track: WARC usage in NAS compared to Archive-it (Tue Hejlskov Larsen)

Coder track: NAS 5.0 code redesign (Mikis Seth Sørensen), Completed and remaining tasks (Søren Vejrup Carlsen)

14:30 - 17:00 Curator track: monitoring and QA crawls with Heritrix 3 (Annick Le Follic and Géraldine Camile)

15:30 - 15:45 Coffee break

17:00 - 17:30 Tour of web archiving activities in Estonia (Jaanus Kõuts)

19:00 - Dinner

Schedule for Day 2 (Friday 30)

Location: *Cupola Hall*

09:00 - 09:30 Harvesting complex websites: experiments with Archive-it 4.9/5.0 using 3.3.0 with Umbra (Tue Hejlskov Larsen)

Experience With IA Umbra (Colin Rosenthal)

09:30 - 11:15 Digging in the data mines of the Net Archive (Per Møldrup-Dalum)

Per presenting a study he just run on DK collections, all presenting on current practices and questions.

Details on the file identification experiment using Nanite: A Weekend With Nanite

Details on the "can we trust the MIME type as it was reported by the web server" experiment: http://rpubs.com/perdalum/de-dup1

Details on the comparison of the domains of the two broadcast companies Comparing the domains of two Danish broadcast companies

The easiest way to get started with R: RStudio

My fork of JWAT-tools for easy extraction of craw.log files: https://bitbucket.org/perdalum/jwat-tools/branch/netarkivet

11:15 - 11:30 Coffee break

11:30 - 13:00 Heritrix tracks sum-up, review of NAS curator roadmap, community next steps (Sara Aubry)

See the workshop conclusions.

13:00 - 14:00 Lunch

14:00 - 15:30 Ebooks/newspapers: deposit or FTP harvesting

Harvesting e-publications in DK – a short status (Tue Hejlskov Larsen)

Printfiles and e-books deposit at the National Library of Estonia (Liina Abner)

Harvesting digital newspapers at the Bibliothèque nationale de France (Géraldine Camile)

15:30 - 15:45 Coffee break

15:45 - 17:00 Open space for an additional topic, individual discussions, workshop closing