EESTI RAHVUS-RAAMATUKOGU

Becoming a Web Archivist:

My 10 Year Journey in the National Library of Estonia

Tiiu Daniel

National Library of Estonia

IIPC Web Archiving Conference,
New Zealand, Wellington

November 13, 2018

You can't connect the dots looking forward; you can only connect them looking backwards (Steve Jobs)

- 2002 – first contact with web preservation topic
  - Bachelor's thesis *Library's Electronic Catalog as Intermediator of Internet Resources (*Information Sciences, Tallinn University)
- 2008 – senior bibliographer of web resources; Collection Development Department, National Library of Estonia
- 2009 – preparations to relaunch the web archiving program

EESTI RAHVUS-RAAMATUKOGU

Legal aspects of web archiving

- 2006–2016 Legal Deposit Act allowed to collect web publications and make them publicly accessible
- since 2017 - new coplitely revised Legal Deposit Copy Act
  - More specific criteria
  - Restricted access, except if permissions from copyright holders were gained
- Open access to government websites by default (according to Public Information Act)

EESTI RAHVUS-RAAMATUKOGU

Selection principles

- Former policy overly library oriented
- 2010 - expert group of people from different memory and research institutions to cover wider range of interests
- 2011 - new selection policy
- 2010-2015 only selective and event/thematic harvesting
- 2016 – first bulk harvesting of Estonian web content

EESTI RAHVUS-RAAMATUKOGU

Aquisition (methods and tools)

2010-2015

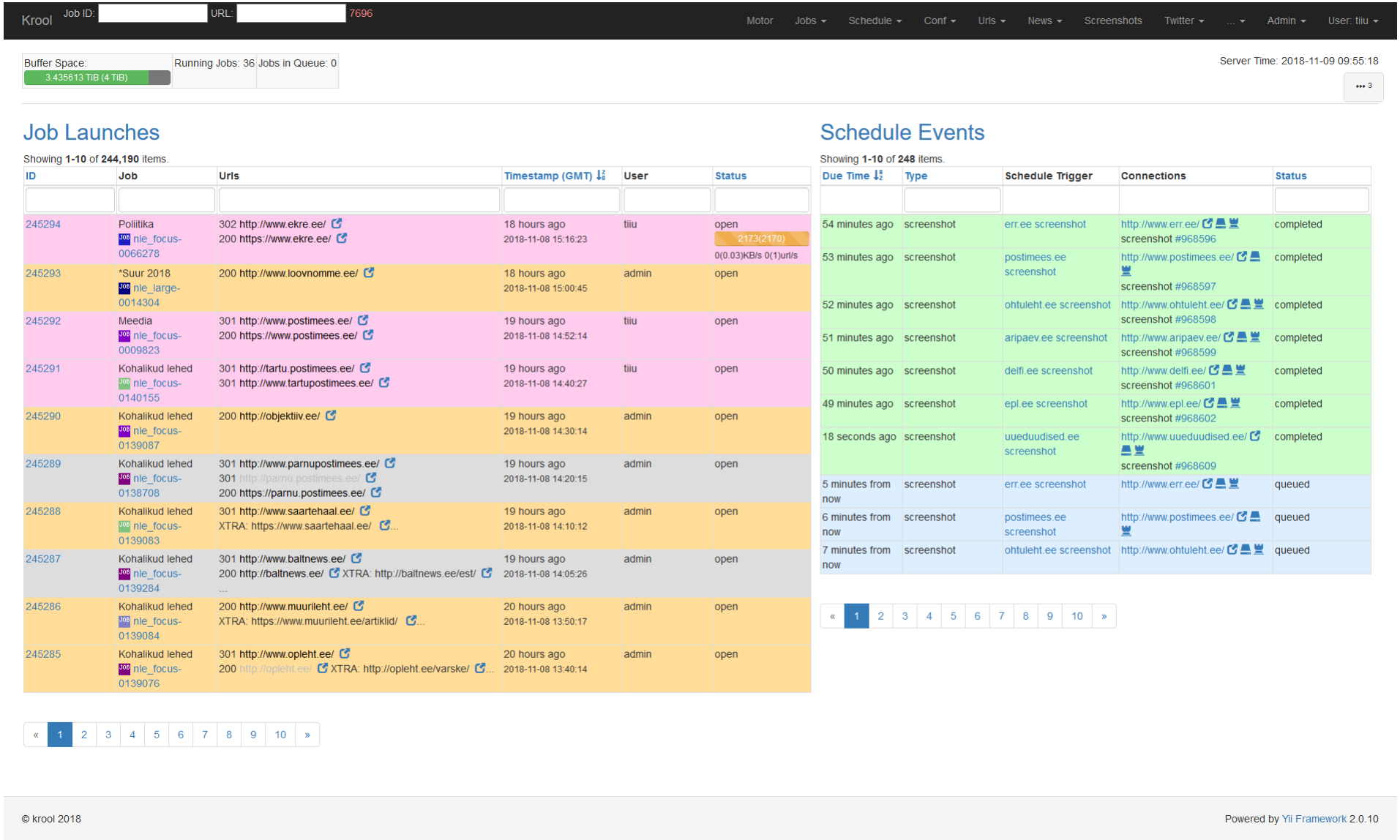
- Selective, event/topical
- Netarcive Suite
- Heritrix version 1.14
- Multiple seeds in one job (topically divided)
- QA – manual browsing

Since 2016

- Selective, event/topical, bulk
- in-house built curation tool
- Heritrix version 3.3
- One seed per job (with few exceptions)
- QA – manual browsing, Heritrix reports analysis
- Screenshots of websites' homepages (since 2016)
- Deduplication (since 2017)

EESTI RAHVUS-RAAMATUKOGU

Curator tool „Krool“



Content description and discovery

- Thematic catalog of archived websites
- Search by URL or words in metadata fields (title, description, seed/related URL)
- Not standarized yet – implementation of Dublin Core and OCLC's „Descriptive Metadata for Web Archiving" planned in 2019
- Wayback Machine's old version in public portal
- Python WayBack – for internal use (QA)
- Full-text search in coming years

Preservation

- Archive size 28,7 TB
- Bit preservation
- 2 copies: library’s disc space + magnetic tape outside the library
- Long-term preservation solution
  - on the analysis stage with as-is process map freshly done, final results are expected in March 2019

EESTI RAHVUS-RAAMATUKOGU

Formation of the web archive team

Profile	2009-2012	2013	2014	2015	2017-
Web curator	1 full-time	1 full-time + 1 part-time	3 full-time	2 full-time	2 full-time
Technical staff	1 part-time	1 part-time	1 part-time	1 full-time	2 full-time

EESTI RAHVUS-RAAMATUKOGU

IIPC – our supporting pillar

- Member since 2012
- Most trusted source of expertise
- Specific community for a very specific task – almost like a secret sect
- Different communication channels, webinars, annual meetings etc.
- Cooperation through collaborative collections
- Not too active in working groups, more keeping eye on them

EESTI RAHVUS-RAAMATUKOGU

Concluding remarks

- Comparing to 2008, landscape has changed remarkably
  - Web archiving is widely recognized activity among memory institutions
  - Most national libraries (at least in Europe) have launched their nations web preservation programs
  - Plenty of (professional) information and events organized
  - Bigger media coverage
- Web is like a jungle –constant struggle with obstacles in capturing the content (robot blocking, limited technical tools, authenticity etc)

EESTI RAHVUS-RAAMATUKOGU

Concluding remarks

Things I love about my work:

- uniqueness
- very versatile in nature
- encourages creativity
- challenging
- enthusiastic and devoted colleagues
- great international community
- ...etc.

EESTI RAHVUS-RAAMATUKOGU

Thank You!

Tiiu Daniel
tiiu.daniel@nlib.ee
http://veebiarhiiv.digar.ee

EESTI
RAHVUS-
RAAMATUKOGU