



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

Legal aspects of web archiving

- ❖ 2006–2016 Legal Deposit Act allowed to collect web publications and make them publicly accessible
- ❖ since 2017 - new completely 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)

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

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)

Curator tool „Krool“

Krool Job ID: URL: 7696

Motor Jobs Schedule Conf Uris News Screenshots Twitter ... Admin User: tiu

Buffer Space: Running Jobs: 36 Jobs in Queue: 0
3.435613 TIB (4 TIB)

Server Time: 2018-11-09 09:55:18

...

Job Launches

Showing 1-10 of 244,190 items.

ID	Job	Uris	Timestamp (GMT) ↓	User	Status
245294	Politika nie_focus-0066278	302 http://www.ekre.ee/ 200 https://www.ekre.ee/	18 hours ago 2018-11-08 15:16:23	tiu	open 2173(2170) 0(0.03)KB/s 0(1)ur/s
245293	*Suur 2018 nie_large-0014304	200 http://www.loovnomme.ee/	18 hours ago 2018-11-08 15:00:45	admin	open
245292	Meedia nie_focus-0009823	301 http://www.postimees.ee/ 200 https://www.postimees.ee/	19 hours ago 2018-11-08 14:52:14	tiu	open
245291	Kohalikud lehed nie_focus-0140155	301 http://tartu.postimees.ee/ 301 http://www.tartupostimees.ee/	19 hours ago 2018-11-08 14:40:27	tiu	open
245290	Kohalikud lehed nie_focus-0139087	200 http://objektiv.ee/	19 hours ago 2018-11-08 14:30:14	admin	open
245289	Kohalikud lehed nie_focus-0138708	301 http://www.parnupostimees.ee/ 301 http://parnu.postimees.ee/ 200 https://parnu.postimees.ee/	19 hours ago 2018-11-08 14:20:15	admin	open
245288	Kohalikud lehed nie_focus-0139083	301 http://www.saartehaal.ee/ XTRA: https://www.saartehaal.ee/ ...	19 hours ago 2018-11-08 14:10:12	admin	open
245287	Kohalikud lehed nie_focus-0139284	301 http://www.baltnews.ee/ 200 http://baltnews.ee/ XTRA: http://baltnews.ee/est/ ...	19 hours ago 2018-11-08 14:05:26	admin	open
245286	Kohalikud lehed nie_focus-0139084	200 http://www.muurieht.ee/ XTRA: https://www.muurieht.ee/artiklid/ ...	20 hours ago 2018-11-08 13:50:17	admin	open
245285	Kohalikud lehed nie_focus-0139076	301 http://www.opleht.ee/ 200 http://opleht.ee/ XTRA: http://opleht.ee/varske/ ...	20 hours ago 2018-11-08 13:40:14	admin	open

« 1 2 3 4 5 6 7 8 9 10 »

Schedule Events

Showing 1-10 of 248 items.

Due Time ↓	Type	Schedule Trigger	Connections	Status
54 minutes ago	screenshot	err.ee screenshot	http://www.err.ee/ screenshot #968596	completed
53 minutes ago	screenshot	postimees.ee screenshot	http://www.postimees.ee/ screenshot #968597	completed
52 minutes ago	screenshot	ohtuleht.ee screenshot	http://www.ohtuleht.ee/ screenshot #968598	completed
51 minutes ago	screenshot	aripaev.ee screenshot	http://www.aripaev.ee/ screenshot #968599	completed
50 minutes ago	screenshot	delfi.ee screenshot	http://www.delfi.ee/ screenshot #968601	completed
49 minutes ago	screenshot	epl.ee screenshot	http://www.epl.ee/ screenshot #968602	completed
18 seconds ago	screenshot	uueduudised.ee screenshot	http://www.uueduudised.ee/ screenshot #968609	completed
5 minutes from now	screenshot	err.ee screenshot	http://www.err.ee/ screenshot #968610	queued
6 minutes from now	screenshot	postimees.ee screenshot	http://www.postimees.ee/ screenshot #968611	queued
7 minutes from now	screenshot	ohtuleht.ee screenshot	http://www.ohtuleht.ee/ screenshot #968612	queued

« 1 2 3 4 5 6 7 8 9 10 »

Content description and discovery

- ❖ Thematic catalog of archived websites
- ❖ Search by URL or words in metadata fields (title, description, seed/related URL)
- ❖ Not standardized 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

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

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

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)

Concluding remarks

Things I love about my work:

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

Thank You!

Tiiu Daniel

tiiu.daniel@nlib.ee

<http://veebiarhiiv.digar.ee>