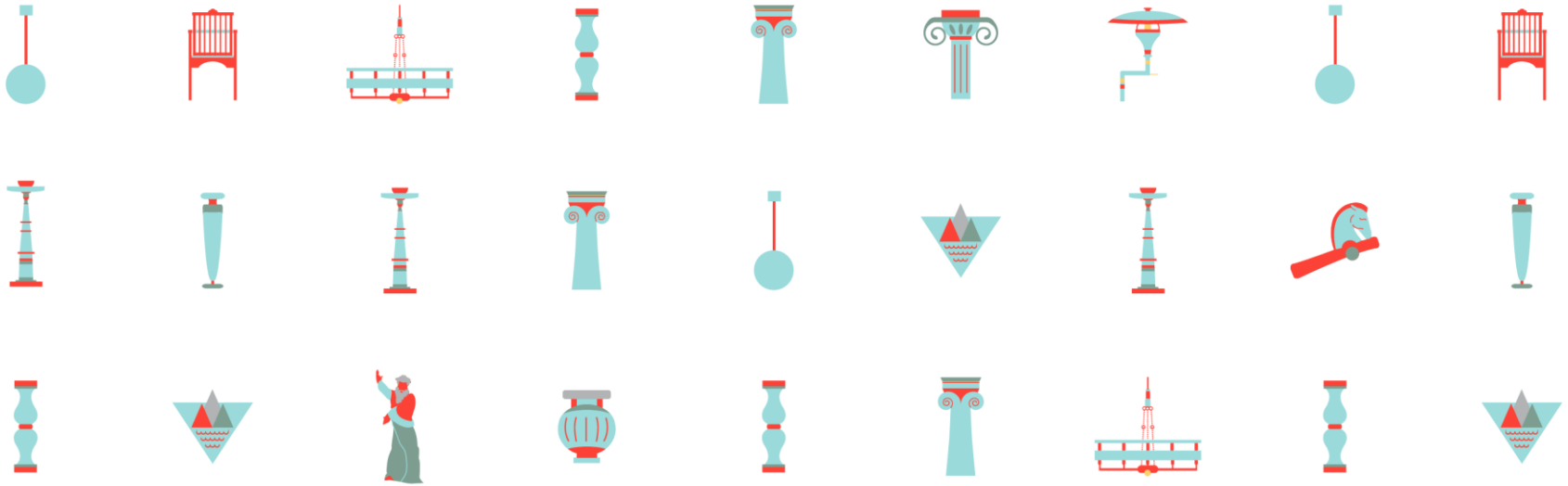
NARODNA IN UNIVERZITETNA KNJIŽNICA

WEB ARCHIVING OVERVIEW

National and University Library - Slovenia



Janko Klasinc | janko.klasinc@nuk.uni-lj.si | +386 01 2001 211

EARLY PROJECTS

NARODNA IN UNIVERZITETNA KNJIŽNICA

2002 – 2004

Slovenian electronic web publications collecting and archiving methodology

2003 – 2004

Development and analysis of slovenian digitized and electronic publications collection of national importance

LEGAL BASIS

NARODNA IN UNIVERZITETNA KNJIŽNICA

2006

Legal deposit law

(Zakon o obveznem izvodu publikacij (Ur. list RS, št. 69/06 in 86/09)

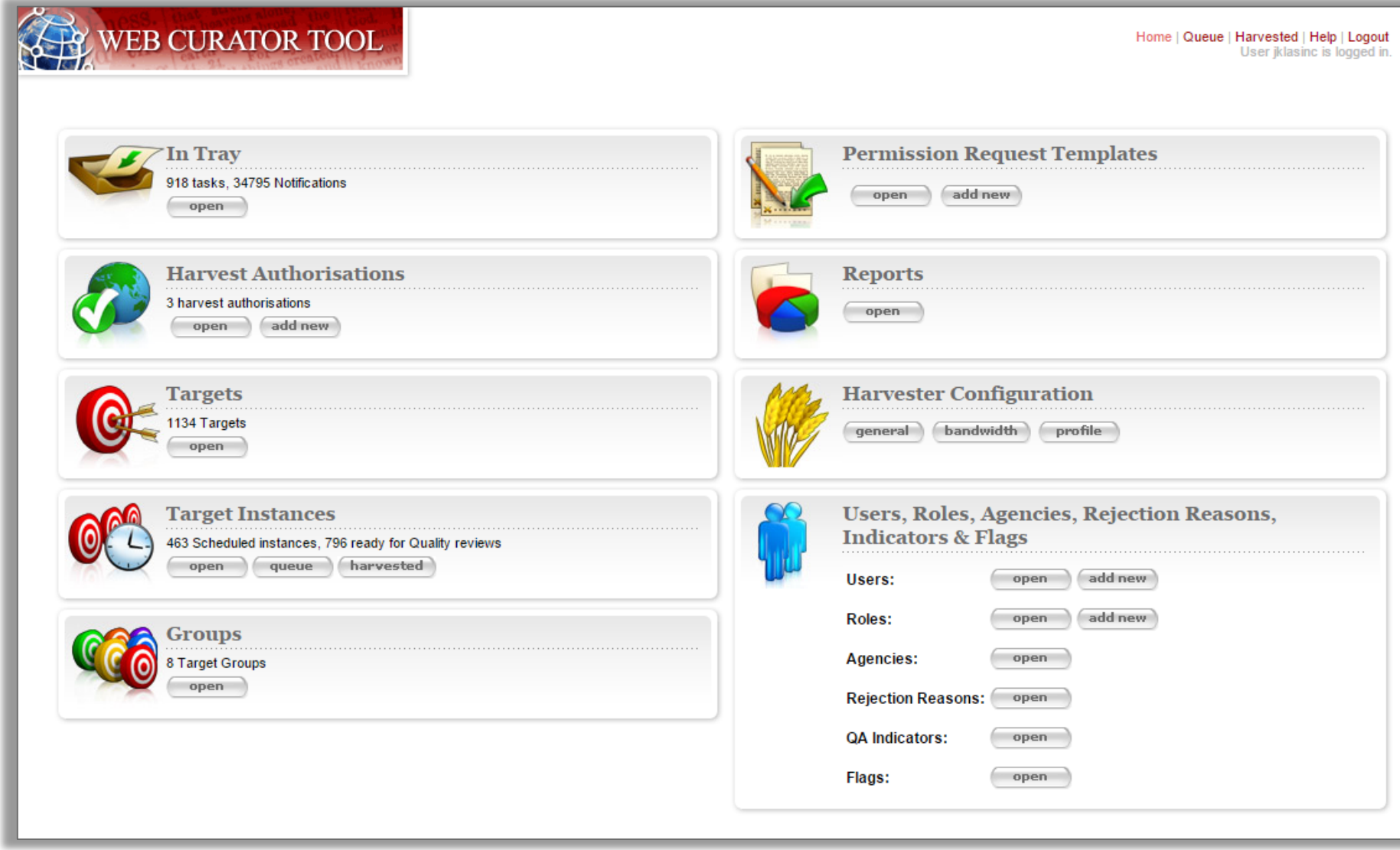
2007

Regulation on types and selection of electronic publications for legal deposit

(*Pravilnik o vrstah in izboru elektronskih publikacij za obvezni izvod, (Ur. list RS, št. 90/07*)

CRAWLING

NARODNA IN UNIVERZITETNA KNJIŽNICA

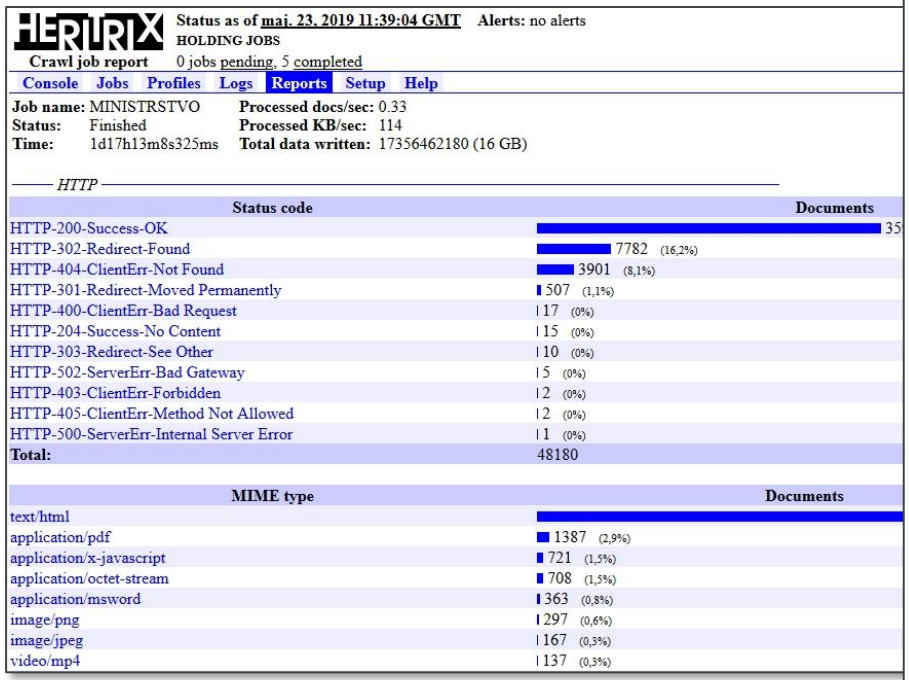


CRAWLING

2014 –

National domain .si crawl (biannually)

Heritrix 1.14.4. and 3.4



HERITRIX

Status as of maj. 23, 2019 11:39:04 GMT Alerts: no alerts

HOLDING JOBS

Crawl job report 0 jobs pending, 5 completed

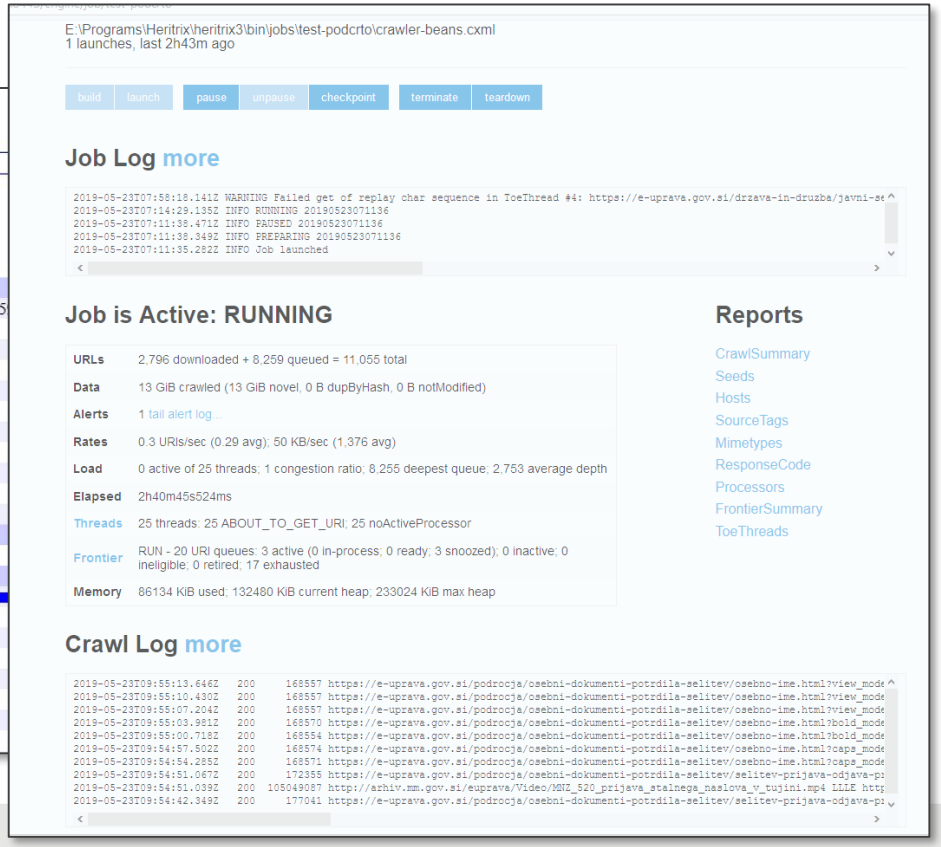
Console Jobs Profiles Logs Reports Setup Help

Job name: MINISTRSTVO **Processed docs/sec:** 0.33
Status: Finished **Processed KB/sec:** 114
Time: 1d17h13m8s325ms **Total data written:** 17356462180 (16 GB)

HTTP

Status code	Documents
HTTP-200-Success-OK	35
HTTP-302-Redirect-Found	7782 (16,2%)
HTTP-404-ClientErr-Not Found	3901 (8,1%)
HTTP-301-Redirect-Moved Permanently	507 (1,1%)
HTTP-400-ClientErr-Bad Request	17 (0%)
HTTP-204-Success-No Content	15 (0%)
HTTP-303-Redirect-See Other	10 (0%)
HTTP-502-ServerErr-Bad Gateway	5 (0%)
HTTP-403-ClientErr-Forbidden	2 (0%)
HTTP-405-ClientErr-Method Not Allowed	2 (0%)
HTTP-500-ServerErr-Internal Server Error	1 (0%)
Total:	48180

MIME type	Documents
text/html	
application/pdf	1387 (2,9%)
application/x-javascript	721 (1,5%)
application/octet-stream	708 (1,5%)
application/msword	363 (0,8%)
image/png	297 (0,6%)
image/jpeg	167 (0,3%)
video/mp4	137 (0,3%)



E:\Programs\Heritrix\heritrix3\bin\jobs\test-podcrto\crawler-beans.cxml
1 launches, last 2h43m ago

build launch pause unpause checkpoint terminate teardown

Job Log more

```
2019-05-23T07:58:18.141Z WARNING Failed get of replay char sequence in ToeThread #4: https://e-uprava.gov.si/drzava-in-druzba/javni-se
2019-05-23T07:14:29.135Z INFO RUNNING 20190523071136
2019-05-23T07:11:38.471Z INFO PAUSED 20190523071136
2019-05-23T07:11:38.349Z INFO PREPARING 20190523071136
2019-05-23T07:11:35.282Z INFO Job launched
```

Job is Active: RUNNING

URLs	2,796 downloaded + 8,259 queued = 11,055 total
Data	13 GiB crawled (13 GiB novel, 0 B dupByHash, 0 B notModified)
Alerts	1 tail alert log...
Rates	0.3 URIs/sec (0.29 avg); 50 KB/sec (1,376 avg)
Load	0 active of 25 threads; 1 congestion ratio; 8,255 deepest queue; 2,753 average depth
Elapsed	2h40m45s524ms
Threads	25 threads: 25 ABOUT_TO_GET_URI; 25 noActiveProcessor
Frontier	RUN - 20 URI queues: 3 active (0 in-process; 0 ready; 3 snoozed); 0 inactive; 0 ineligible; 0 retired; 17 exhausted
Memory	86134 KiB used; 132480 KiB current heap; 233024 KiB max heap

Reports

CrawlSummary
Seeds
Hosts
SourceTags
Mimetypes
ResponseCode
Processors
FrontierSummary
ToeThreads

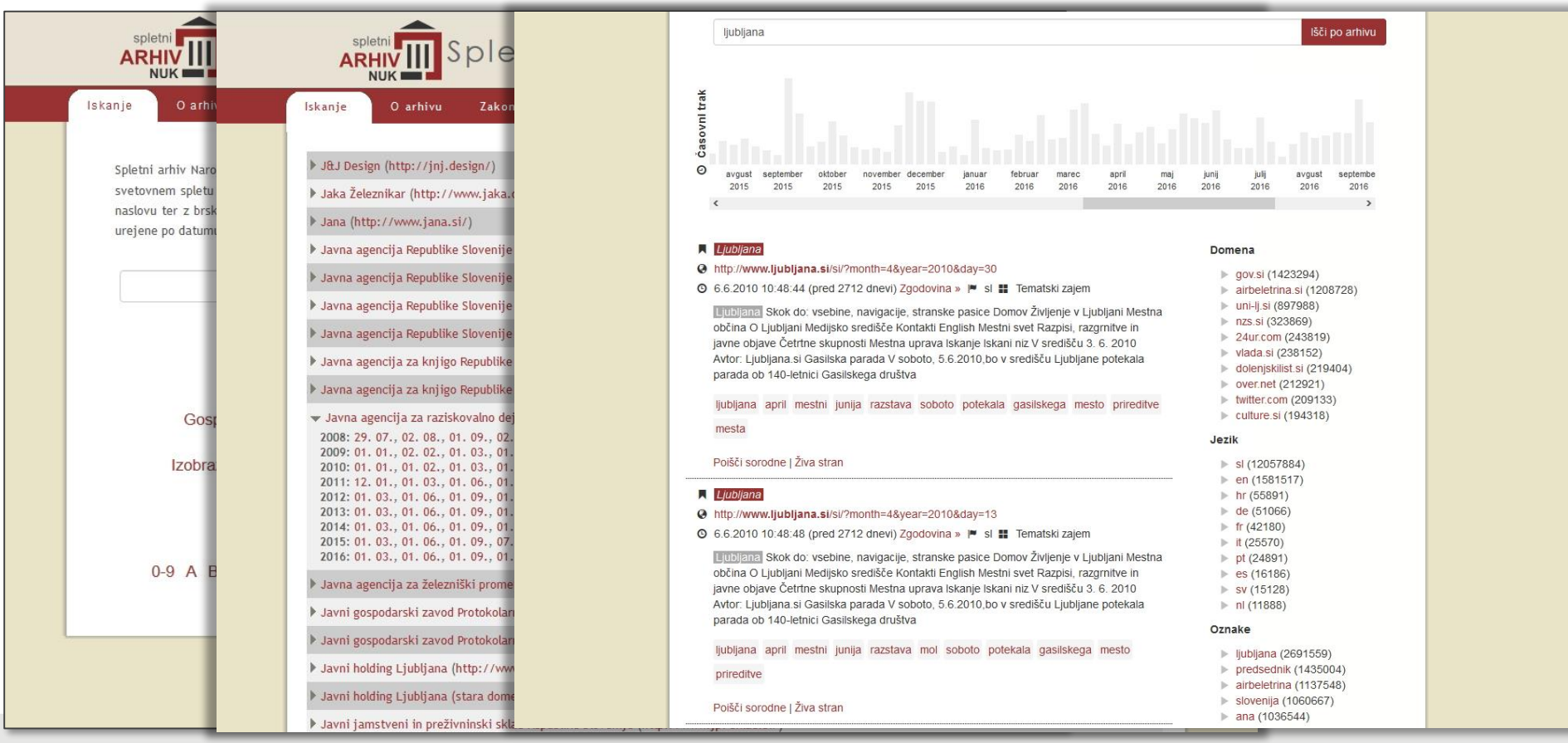
Crawl Log more

```
2019-05-23T09:55:13.646Z   200     168557 https://e-uprava.gov.si/podrocja/osebni-dokumenti-potrdila-selitev/osebno-ime.html?view_mode
2019-05-23T09:55:10.430Z   200     168557 https://e-uprava.gov.si/podrocja/osebni-dokumenti-potrdila-selitev/osebno-ime.html?view_mode
2019-05-23T09:55:07.204Z   200     168557 https://e-uprava.gov.si/podrocja/osebni-dokumenti-potrdila-selitev/osebno-ime.html?view_mode
2019-05-23T09:55:03.981Z   200     168570 https://e-uprava.gov.si/podrocja/osebni-dokumenti-potrdila-selitev/osebno-ime.html?bold_mode
2019-05-23T09:55:00.718Z   200     168554 https://e-uprava.gov.si/podrocja/osebni-dokumenti-potrdila-selitev/osebno-ime.html?bold_mode
2019-05-23T09:54:57.502Z   200     168574 https://e-uprava.gov.si/podrocja/osebni-dokumenti-potrdila-selitev/osebno-ime.html?caps_mode
2019-05-23T09:54:54.285Z   200     168571 https://e-uprava.gov.si/podrocja/osebni-dokumenti-potrdila-selitev/osebno-ime.html?caps_mode
2019-05-23T09:54:51.067Z   200     172355 https://e-uprava.gov.si/podrocja/osebni-dokumenti-potrdila-selitev/selitev-prijava-odjava-p
2019-05-23T09:54:51.039Z   200  105049087 http://arhiv.mm.gov.si/euprava/Video/MNZ_520_prijava_stalnega_naslova_v_tujini.mp4 LLLE htt
2019-05-23T09:54:42.349Z   200     177041 https://e-uprava.gov.si/podrocja/osebni-dokumenti-potrdila-selitev/selitev-prijava-odjava-p
```

ACCESS

2011 -

Wayback Machine



DATA COLLECTED

National domain, selective & thematic crawls:

- 560.066 domains
- 513.793.472 URLs
- 45,5 TB

Saff:

0,25 FTE?

CURRENT ACTIVITIES

NARODNA IN UNIVERZITETNA KNJIŽNICA

- moving WCT, Heritrix & Wayback to new servers (separating crawling from access);
- focused crawl of 50 government domains before the content is moved to a single domain;
- providing access to the national domain crawls;
- rethinking legal basis for free access.

NARODNA IN UNIVERZITETNA KNJIŽNICA

THANK YOU!

janko.klasinc@nuk.uni-lj.si
+386 01 2001 211