Submitted on: 1.7.2015

First crawling of the Slovenian National web domain *.si: pitfalls, obstacles and challenges

Matjaž Kragelj
Head, Digital Library Development Department and Head, Information Technology and Digital Library Division, National and University Library, Slovenia
(matjaz.kragelj@nuk.uni-lj.si)

Mitja Kovačič
Digital Library Development Department, Information Technology and Digital Library Division, National and University Library, Slovenia
(mitja.kovacic@nuk.uni-lj.si)

Abstract:

The National and University Library (NUK) has been archiving the web for almost fifteen years. During the last six years, we have been trying to act on different levels of harvesting. For most of the time, we have dealt with harvesting of selected web sites that might be significant for future generations. The harvesting process runs smoothly, with the exception of some technical difficulties resulting from the use of scripted languages (for instance Ajax, Flash, Java script, asynchronous transmissions, real time streaming protocols, etc.). The number of archived web pages keeps growing very fast. We are also very successful in harvesting social media web sites with tools developed in NUK.

Being aware that the amount of the web pages cannot be compared with the harvested one - it is much more extensive – we decided to start the Slovenian domain (.si) harvesting.*

The first domain harvesting was successful; however, we realized that much deeper and broader levels should be harvested by using heuristic methods.

*Our experiences showed that most informative web contents are hidden beneath the *.si domain's data provided by ARNES (Academic Research Network of Slovenia), therefore, the contents are not accessible.*

The paper presents the results of the first harvesting iteration of the Slovenian web. Further, on a sample of the first hundred domains, the results of the first and second harvesting iteration will be compared and analysed. At the end, the relevance of data acquired in the harvested web pages as a digital library complementary data source will be presented.

Keywords: web archiving, harvesting, national domain, social networks harvesting, digital library

For more than a decade, the National and University Library (NUK) has invested efforts in archiving the publicly accessible contents on the Slovenian web. Since its origins in the 90's, the web has exponentially grown in all its dimensions - regarding the number of users[1], as well as regarding accessible links and data on the web[2] and the traffic produced on this media[3]. Due to a drastic increase of information on the web, each attempt to preserve all contents published on the web seems impossible. This is the reason that only a few organizations in the world, mainly legal deposit organizations, universities and research institutions or national libraries, are prepared to undertake this challenge. Being aware of the difficulties of safely storing and preserving the web, most of them prefer to focus on smaller web domain sets or specific topics, which are manageable in real time, and they try to preserve only part of the whole information mosaic available on the web.

In Slovenia, web archiving is regulated by the Legal Deposit Law, passed in 2006 (Official Gazette, 69/06). According to the Article nr. 2 of this Law, the publications on the web are defined as "electronic books, electronic journals and newspapers which are accessible on the web and other similar publications" and according to the Article nr. 4 these are legal deposit copies.

In 2007, the Slovenian government adopted the" Regulations on types and selection criteria for legal deposit of electronic publications" ("Pravilnik o vrstah in izboru elektronskih publikacij za obvezni izvod") (Official Gazette, nr. 90/07), which are a supplement to the Legal Deposit Law. In the Regulations, in addition to books, journals, newspapers and articles published on the web, as web publications are also listed "web sites (pages) of organizations, people, events, information portals, web services, databases, web news, web conferences (forums), newsletters and other different electronic contents like video and audio records, interactive maps, city maps, software, computer games, web art, blogs, wikies, e-learning and similar".

The Law assigns NUK as the organization responsible for acquiring the electronic publications published on the web by web searching and harvesting. Should this not be possible, the depositors themselves have to deliver their contents published on the web to NUK.

The main selection criteria for archiving the Slovenian cultural and scientific heritage on the web are: contents with the Slovenian authorship, web sites on Slovenia and web sites in the Slovenian language.

[1] Hong, Kaylene (2014). Akamai: Global average web speed up 24% annually to 3.9 Mbps, 20% of connections now above 10 Mbps. http://thenextweb.com/insider/2014/06/27/akamai-global-average-internet-speed-24-year-year-3-9-mbps-mobile/

[2] Brown, Jonathon (2013). Internet Growth Timeline Charts. Available on 4 June from page: http://www.trendhunter.com/trends/history-of-the-internet

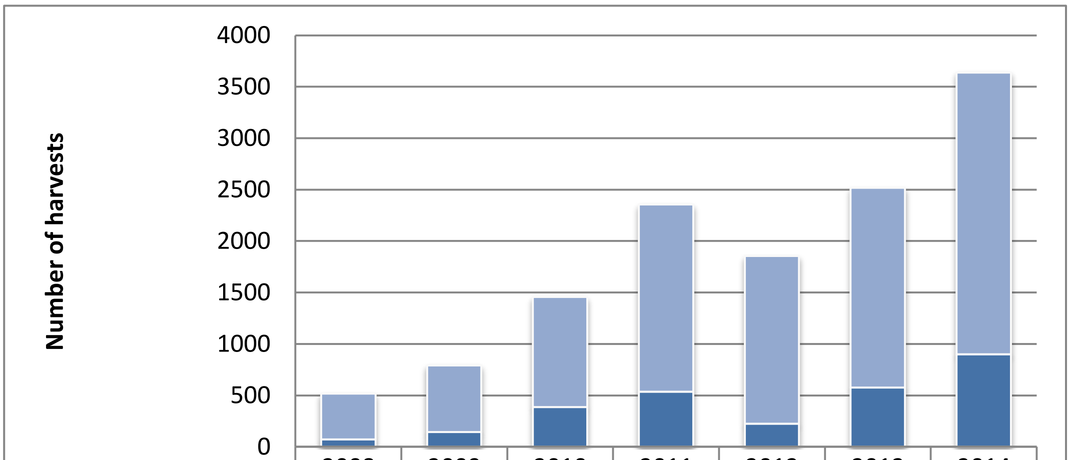
[3] Beyster, J. Robert & Daniels, Mike (2013). Surviving Exponential Growth: Lessons from Network Solutions. Available on 4 June 2015 from page: http://www.xconomy.com/san-diego/2013/11/07/surviving-exponential-growth-lessons-network-solutions/

In NUK, we manage digital preservation processes of web contents in three parallel levels:

- Slovenian selective domain harvesting
- Twitter social network harvesting
- Annual iterative harvesting of the Slovenian domain *.si

1. Slovenian selective domain harvesting

Based on the Legal Deposit Law and aforementioned Regulations, NUK has periodically harvested and preserved the selected web domains, which represent important part of the Slovenian cultural heritage on the World Wide Web since 2008. These domains are categorized in 12 different topics, i.e. society, economy and industry, human sciences, education and research, public media, nature and environment, science and technology, leisure, tourism, travel, sports and recreation, arts and culture, government, politics, law, health and medical sciences. The Graph 1 shows the number of (successful and failed) harvests undertaken during the last seven years. Before 2008, NUK carried out several experiments and trials with web harvesters and different harvesting software[4].

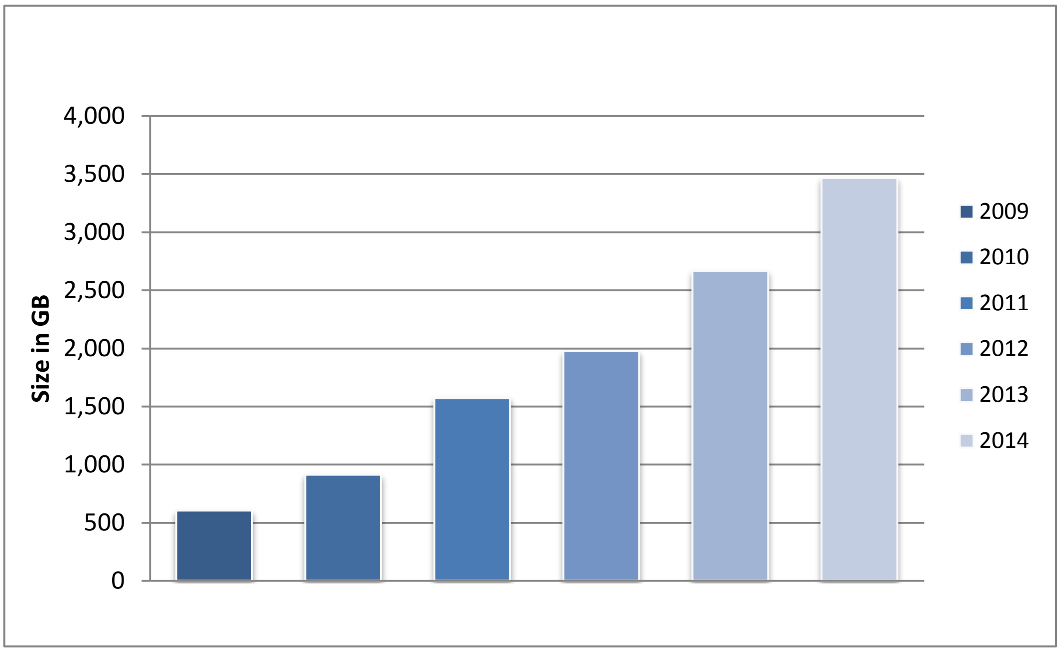


	2008	2009	2010	2011	2012	2013	2014
■ Successful harvests	446	649	1071	1821	1629	1940	2738
■ Failed harvests	73	144	386	535	226	577	901

Graph 1: Number of successful and failed harvests from 2008 to 2014

The Graph 2 shows the increase of harvested contents since 2009. In 2014, **3.5 TB** of web sites were successfully harvested, which represents a **30% increase** compared to 2013. In addition to periodic Slovenian web sites harvesting, we also do selective harvesting on the topics mentioned before. In this way, we respond to the situation in the country and broader region in order to preserve the whole information on the events of national significance (for instance, the elections in Slovenia, Winter Olympic Games in Sochi, etc.).

[4] Kavčič-Čolić, A., Grobelnik, M. (2004). Archiving the Slovenian web : recent experiences. In: Proceedings : 4th international web archiving workshop (IWAW04), held in conjunction with the 8th European conference on research and advanced technologies for digital libraries, september 16 2004, Bath, UK. - Bath : University Available on 5 June 2015 at web page: http://www.iwaw.net/04/index.html.



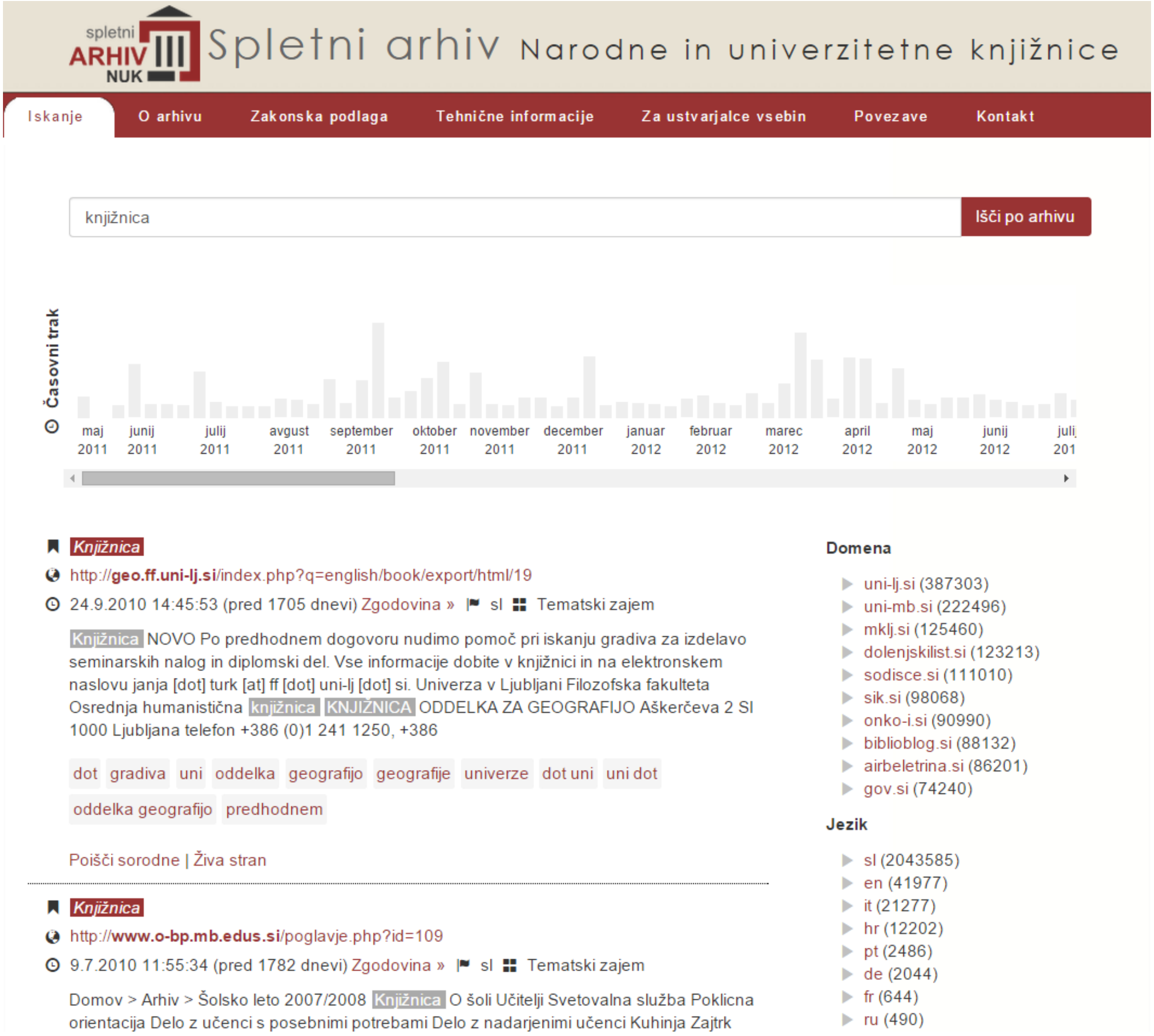
Graph 2: Increase of harvested contents in GB from 2009 to 2014

In May 2015, selective harvesting is focused on 1250 web domains. In the NUK web archive, we have more than 100.000.000 web pages on different topics. These web pages can be full-text searched; the index on SOLR[5] platform is used for indexing.

Figure 1 shows the user interface of the NUK web archive. The users have the possibility for searching through all the contents in the web archive. There is also a timeline, which gives the information on the number of harvests, number of domains harvested and available versions of the web page, which could be accessed. The user can filter the search results by language, while SOLR index provides part of the text taken from the web page in the result list. Among other SOLR platform functions we also use "More Like This", which proposes and finds similar web sites to the one found based on similarity analysis.

In the actual version of the web archive user interface, we usually build vocabularies based on the frequently used words and terms in a web page, which serve as a filter for new searches. In 2015, we plan to implement Natural Language Processing[6] (NLP) algorithms for optimal extraction of key words from web pages. Assigning entities in web pages will enable sorting, classification of web sites in clusters according to the main topic of the texts.

[5] Solr, http://lucene.apache.org/solr/, access May 2015
[6] Natural Language Processing, http://stackoverflow.com/tags/nlp/info, access May 2015



Knjižnica
http://geo.ff.uni-lj.si/index.php?q=english/book/export/html/19
24.9.2010 14:45:53 (pred 1705 dnevi) Zgodovina » sl Tematski zajem

Knjižnica NOVO Po predhodnem dogovoru nudimo pomoč pri iskanju gradiva za izdelavo seminarskih nalog in diplomski del. Vse informacije dobite v knjižnici in na elektronskem naslovu janja [dot] turk [at] ff [dot] uni-lj [dot] si. Univerza v Ljubljani Filozofska fakulteta Osrednja humanistična knjižnica KNJIŽNICA ODDELKA ZA GEOGRAFIJO Aškerčeva 2 SI 1000 Ljubljana telefon +386 (0)1 241 1250, +386

dot gradiva uni oddelka geografijo geografije univerze dot uni uni dot oddelka geografijo predhodnem

Poišči sorodne | Živa stran

Knjižnica
http://www.o-bp.mb.edus.si/poglavje.php?id=109
9.7.2010 11:55:34 (pred 1782 dnevi) Zgodovina » sl Tematski zajem

Domov > Arhiv > Šolsko leto 2007/2008 Knjižnica O šoli Učitelji Svetovalna služba Poklicna orientacija Delo z učenci s posebnimi potrebami Delo z nadarjenimi učenci Kuhinja Zajtrk

Domena

- uni-lj.si (387303)
- uni-mb.si (222496)
- mklj.si (125460)
- dolenjskilist.si (123213)
- sodisce.si (111010)
- sik.si (98068)
- onko-i.si (90990)
- biblioblog.si (88132)
- airbeletrina.si (86201)
- gov.si (74240)

Jezik

- sl (2043585)
- en (41977)
- it (21277)
- hr (12202)
- pt (2486)
- de (2044)
- fr (644)
- ru (490)

Figure 1: Web archive user interface and search results

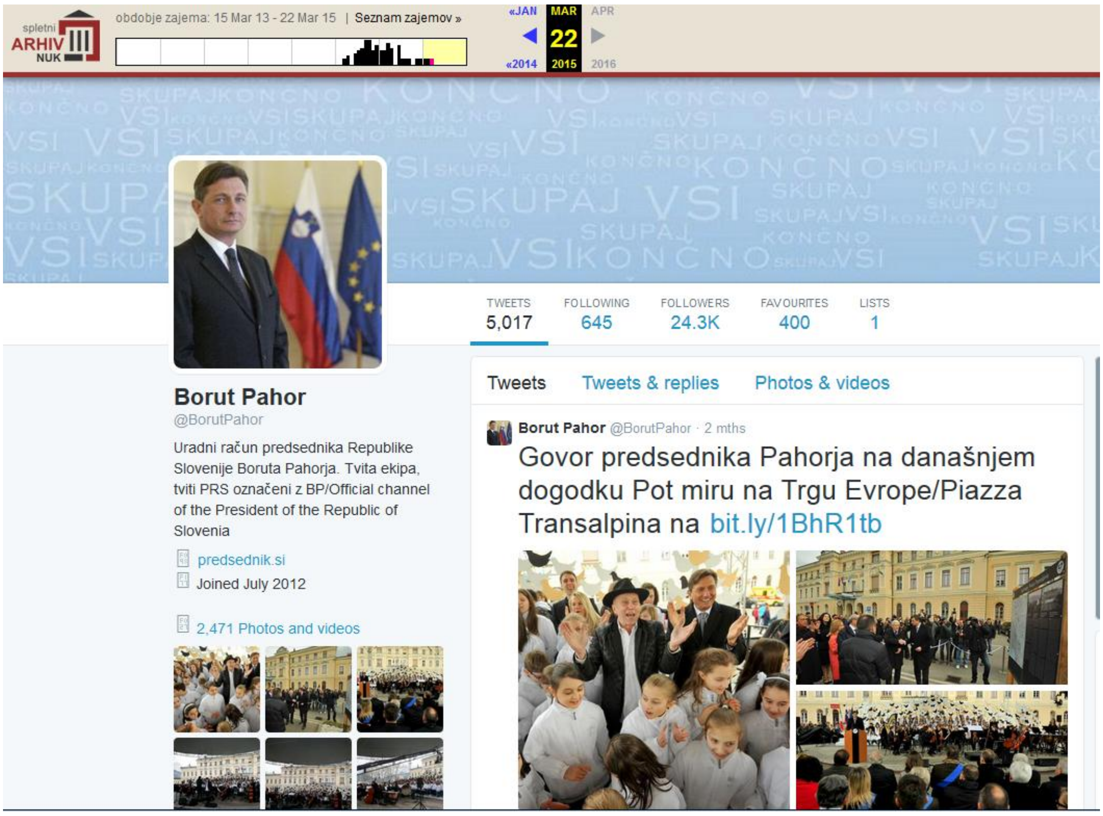
2. Social networks harvesting

In the last few years, the number of users of social networks, like Facebook, Twitter, Google+, Instagram, and others [7] is rapidly increasing. By using these tools, mostly the young people spread their thoughts, information about events and other information directly in real time. The knowledge on and use of web technologies and tools as well as the acquaintance with scripted languages is not necessary any more – on the contrary, the user needs just a few seconds to download her/his photo, to comment on some event or publish any text on the web. To achieve this kind of communication between the user and the system and to make an optimal end-user experience, it was necessary to adjust and upgrade the web technologies. AJAX technologies (Asynchronous JavaScript and XML)[8], which provide the possibility of asynchronous exchange of data between the client (user) and the server, appeared on the web pages.

Since this kind of web pages need user feedback for its functioning (they are interactive), their harvesting is troublesome. Web harvesters do not simulate the functioning of the user

[7] Seznam socialnih omrežij, http://en.wikipedia.org/wiki/List_of_social_networking_websites, access May 2015
[8] Ajax, http://en.wikipedia.org/wiki/Ajax_%28programming%29, access May 2015

interface (browser), the content, which should be dynamically uploaded upon user request, cannot be generated and as a consequence it is not uploaded. Therefore, it was necessary to develop tools, which can simulate the functioning of the browser, and thus, upload all the desired content. With the use of the product PhantomJS[9] (scripted, headless browser used for automating web page interaction), it is possible to harvest this kind of web pages, which require user interaction for presenting the contents. Twitter[10] is an example of this kind of interactive web pages. With a simple harvesting, it is possible to get the first page only (the first few twits). On the other hand, with the help of PhantomJS and some additional software codes, it is possible to harvest the complete user's Twitter profile and to store it for further access. An example is the archived Twitter of the Slovenian President Borut Pahor in Figure2.



Borut Pahor
@BorutPahor

Uradni račun predsednika Republike Slovenije Boruta Pahorja. Tvita ekipa, tviti PRS označeni z BP/Official channel of the President of the Republic of Slovenia

predsednik.si
Joined July 2012

2,471 Photos and videos

Borut Pahor @BorutPahor · 2 mths
Govor predsednika Pahorja na današnjem dogodku Pot miru na Trgu Evrope/Piazza Transalpina na bit.ly/1BhR1tb

Figure 2: Borut Pahor's twitter in the NUK web archive (http://nukrobi2.nuk.unilj.si:8080/wayback/20150322050023/http://www.twitter.com/BorutPahor/)

In the NUK web archive, we keep twitters from famous Slovenian people in sports, politics and media.

3. Slovenian web domain *.si harvesting

In spring 2014, NUK received the list of web domains from the Academic Research Network of Slovenia (ARNES) [11] register. ARNES builds, maintains and manages the infrastructure,

[9] PhantomJS, http://phantomjs.org/, access May 2015
[10] Twitter, https://twitter.com/, access May 2015
[11] ARNES, http://www.arnes.si/zavod-arnes/predstavitev.html, access May 2015

which connects universities, institutes, research laboratories, museums, schools, databases and digital libraries. The ARNES database comprises of over 100.000 web registered domains. The first harvesting iteration based on *.si domain took place from spring 2014 to spring 2015. In this occasion, we divided the list of domains in equal shares through the whole year in order to accordingly distribute the needed IT resources for this task. The harvesting took part according to the principle “first in the list – first processed”, since the expiring period of the domains in the list was not known.

The harvesting was focused to the third level depth starting from the home page. We set the limit of 500 MB and 10.000 URLs for each web site. The final results of the first harvesting iteration was:

- **DocumentCrawled : 52.219.795**
- **SeedsCrawled : 85.951**
- **RawDataSize : 2.653.099.023.271**
- **Hosts count: 209.225**
- **Time spent: 176d 14h 45m 8s 575ms**

As shown in the results, the harvesting lasted 176 days. We have to emphasize that the works took part consecutive. We collected more than 52 million of documents and from approximately 100,000 domains we managed to search and harvest 85,951 domains. The main reason that we missed to harvest some domains was that they were empty (without contents) or there was a link to other web site outside .si scope (redirection). Since we followed the criteria of “one jump only”, we visited only 209,225 domains from the harvested domains (85,951). Among them were the domains which were accessed through links on the harvested domain (for instance, the links to YouTube, different applications for statistics, links to social networks like Facebook and Twitter, links to topic related pages and organizations, partners, etc.). The total amount of content was over 2.6 TB of data.

The main findings and direction for the next harvesting iteration (2015-2016) are:

The first harvesting iteration of the Slovenian web domain *.si did not produce important technical difficulties. The main finding was that the larger quantity of the relevant content is not available.

Because the ARNES database of registered Slovenian domains contains top domains only (e.g. ***uni-lj.si, gov.si***), most of the contents remain hidden. The analysis of the contents of these domains shows that there is a large number of subdomains.

GOV.SI

The owner of the domain gov.si is the Government of the Republic of Slovenia and it comprises all the government bodies. The web domain http://gov.si contains 66 subdomains on the first two levels. Among the most frequent subdomains are:

SUBDOMAIN	**number of occurrences in the domain**
http://e-uprava.gov.si/	398
http://www.gov.si/	317
http://www.upravneenote.gov.si/	116
http://evem.gov.si/	106

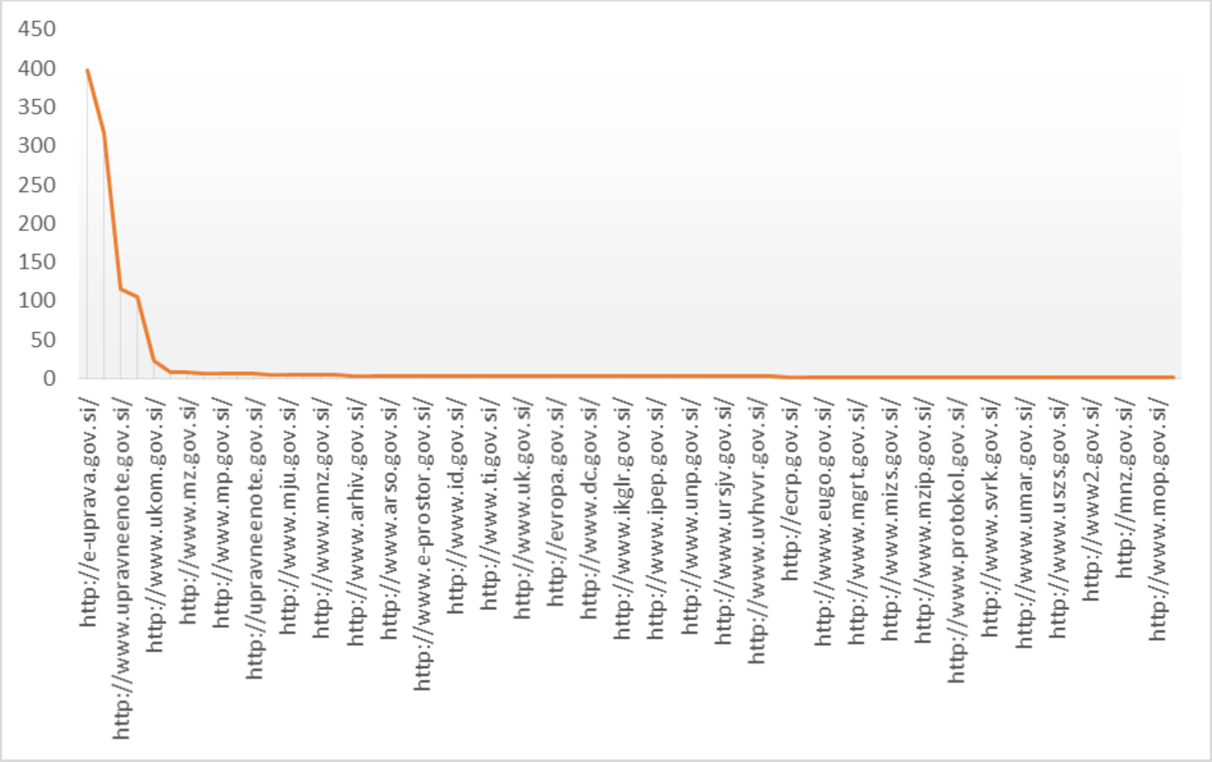


Figure 3: Subdomain occurrences under the domain http://gov.si

Figure 3 shows a cut of web subdomains which appear more frequently as a link in the web domain http://gov.si. The total number of subdomains is 1160. According to the frequency of links on the subdomains and the number of subdomains, we have realized that there is a huge hidden space not accessed by our harvester. In the second harvesting iteration planned in 2015, we will include these subdomains we found in the analysis of the aforementioned top domain in the URLs seed list.

UNI-LJ.SI

A similar case was the analysis of the top domain ***uni-lj.si*** (Figure 4). The owner of this top domain is the University of Ljubljana. It was expected to find in this domain the links to the faculties and other members of the University. Due to the huge gap between the top domain and subdomains, the logarithmic scale is shown on Figure 4.

SUBDOMAINS	**number of occurrences in the domain**
http://www.uni-lj.si/	345571
http://kc.uni-lj.si/	24641
http://www.vpis.uni-lj.si/	1732
http://intranet.uni-lj.si/	1648
http://www.iri.uni-lj.si/	1644
http://www.ff.uni-lj.si/	45
http://www.bf.uni-lj.si/	41
http://www.ntf.uni-lj.si/	37
http://www2.zf.uni-lj.si/	37
http://www3.fgg.uni-lj.si/	36
http://www.fu.uni-lj.si/	29
http://www.aluo.uni-lj.si/	27
http://www.ffa.uni-lj.si/	27
http://www.ef.uni-lj.si/	25
http://www.fe.uni-lj.si/	25

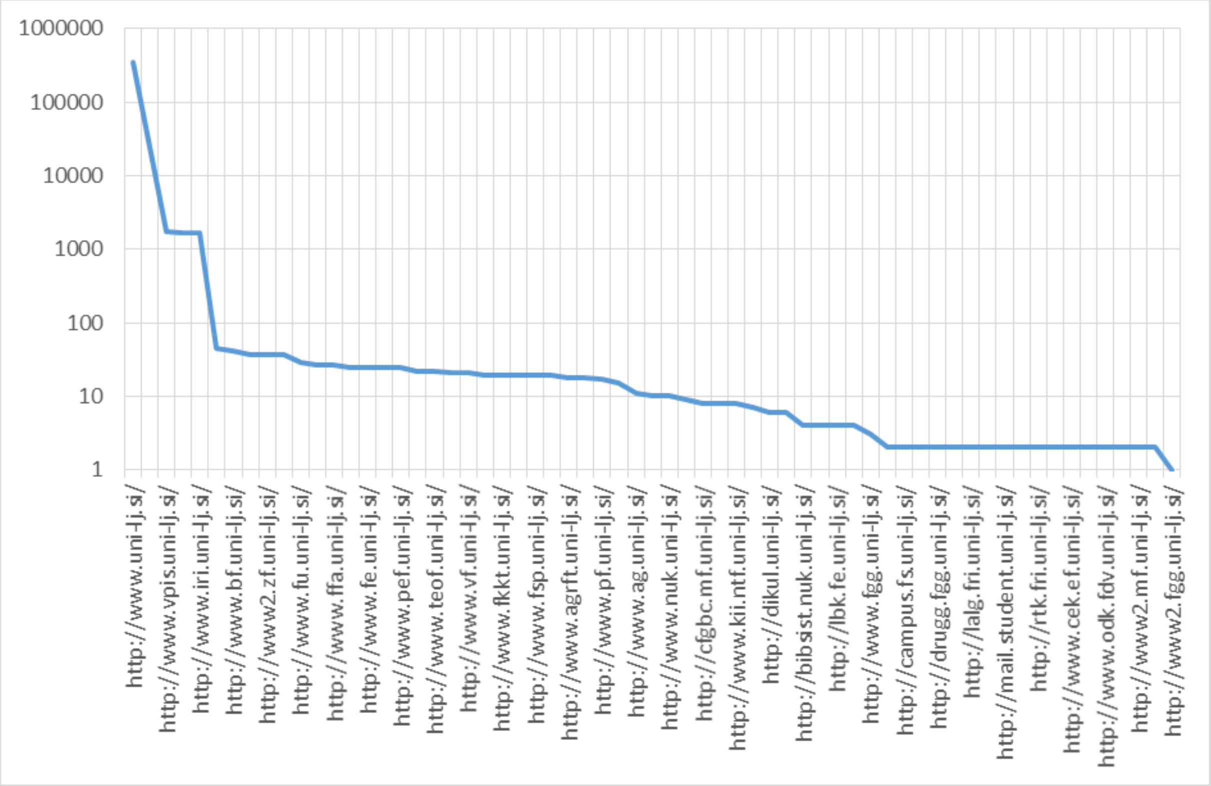


Figure 4: Subdomain occurrences under the top domain **uni-lj.si**

IN CONCLUSION

The web archive will become important over the next years. It represents a collection of contents not any more accessible on the live web and not available in other forms (e.g. the web site of the deceased alpinist Tomaž Humar). The archive offers us a possibility to view the structure and form of some electronic medium in the past. We foresee that for the researchers and other users the value of the web archive will increase in the next decades. In addition to their structure and form, the web pages can provide us the communication spirit over the time. By archiving social networks, we can follow people’s feedbacks to different cultural, sport and other social and natural events.

We expect that the second harvest iteration of the *.si domain, where we will try to harvest the biggest Slovenian top domains, will bring us an exponential growth of data. Based on the top domains examples (uni-lj.si and gov.si), we will be able to access a large quantity of information, which were not collected in the first harvest. In 2015-2016, we will analyse the harvested web sites and their contents (web pages) in order to classify the webpages according to UDC (Universal Decimal Classification). This will help us to link millions of pages according to their content and theme to be accessed as additional collections through the Digital Library of Slovenia. In addition to the collections in the Digital Library of Slovenia, we would like to offer our users the related web sites from the Slovenian web archive.