Submitted on: 8/10/2015

A decade of web archiving in the National and University Library in Zagreb

Karolina Holub (kholub @nsk.hr)
Ingeborg Rudomino (irudomino@nsk.hr)
Croatian Web Archive, National and University Library in Zagreb, Croatia

Abstract:

Due to the dynamic nature of the web, its explosive growth, short lifespan, instability and similar characteristics, the importance of its archiving has become priceless for future generations. The National and University Library in Zagreb (Nacionalna i sveučilišna knjižnica u Zagrebu, NSK), as a memory institution responsible for collecting, cataloguing, archiving and providing access to all types of resources, recognized the significance of collecting and storing online content as part of the NSK's core activities. This is supported by positive legal environment since 1997 when Croatia passed the Law on libraries which subjected online publications to legal deposit. In 2004 NSK established the Croatian Web Archive (Hrvatski arhiv weba, HAW) in collaboration with the University Computing Centre (Srce) and developed a system for capturing and archiving Croatian web resources. From 2004 to 2010 only selective archiving of web resources was conducted according to pre-established selection criteria. Taking into account NSK's responsibility to preserve resources on Croatian social, scientific and cultural history, the importance of taking a snapshot of all publicly available resources under the national top level domain (.hr) was been recognized in 2011. Since then national domain harvestings have been conducted annually. In addition, in 2011 NSK started to run thematic harvestings of national importance.

The paper will present the NSK's ten years' experience in managing web resources with the emphasis on implementation of the system for selective and domain harvesting as well as the challenges for providing access to archived resources. Also, the harvested data from 2004 to 2014 will be analysed. The findings will illustrate the variability of URLs, frequency of harvesting and types of content. The data from the last four .hr harvestings will also be presented.

Keywords: web archiving, selective harvesting, national domain harvesting, Croatian Web Archive, thematic harvesting, legal deposit

Introduction

Internet appearance in Croatia dates to early 1990s when the first internet service provider - CARNet (The Croatian Academic and Research Network)[1] started working and a few years later, in 1993, started administering the national domain (.hr). In 1997 the first Croatian news portal *Internet monitor* appeared. Dynamic nature and growth of the internet, as well as a short lifespan characterizes the content published on this media.

The institution responsible for collecting, building and organizing the national collection of library resources in Croatia, the National and University Library in Zagreb has been receiving legal deposit since 1816. The new Library Law entered into force with new provisions and the law of 1997 recognized the obligation of processing and keeping, the then new content, online publications. The NSK started cataloguing online content in 1998 after the Law was passed and up until 2003 783 resources have been catalogued.[2] Unfortunately, during that period, owing to financial and technical difficulties, and inadequate infrastructure the ingest and storage of these type of resources was not done. This resulted in an irreversible loss of significant part of web content. NSK begun developing a system for capturing and archiving legal deposit of online resources with the University of Zagreb University Computing Centre (Srce)[3] in 2003 which resulted in the establishment of a service called the Digital Archive of Croatian Online Publications (DAMP) in November 2004. In 2010 the service changed its name to Hrvatski arhiv weba (Croatian Web Archive, HAW).

Fig. 1. Croatian Web Archive's homepage

[1] CARNet. Available at: http://www.carnet.hr/

[2] Willer, Mirna; Milinović, Miroslav. Prema trećoj generaciji knjižnično-informacijskih sustava : hibridna knjižnica za hibridne usluge // 8. seminar Arhivi, knjižnice, muzeji : mogućnosti suradnje u okruženju globalne informacijske infrastrukture : zbornik radova. Zagreb : Hrvatsko knjižničarsko društvo, 2005., str. 47.

[3] University of Zagreb University Computing Centre. Available at: http://www.srce.unizg.hr/

HAW is a system for capturing Croatian online resources in order to preserve the authenticity, form and functionality of the archived content. In the beginning, the main approach was to archive only those resources which were bibliographically described, meaning that they were treated as all other library content. [4] Owners and/or content providers[5] of all types of online resources are legally bound to deliver and archive their resources. Today, after 10 years in managing Croatian web resources, it is clear that the system has been following trends in online publishing in Croatia, which began during the second half of the 1990s. Croatian online publishing has caught up with the rest of the world around the year 2000, when massive amounts of online content started to appear, website owners and/or content providers realized the advantages of this kind of publishing. Interestingly enough online publishing was increasing up until 1999 there were only 6% of registered Internet users. [6]

Web harvesting approaches

Selective harvesting

Resources available online are considered published and modern technology allows anyone who posts online to be an author and/or a publisher. Taking into consideration the variety of web content but also the NSK's obligation, as a national bibliographical centre, to collect, build and organize the national collection of online content, the web harvesting started selectively. For that reason the NSK established *Selection criteria* for cataloguing and archiving web resources.[7] The general criteria applied to the traditional publications are identical to those applied to online resources:

- works by Croatian authors published in Croatia and abroad,
- works about Croatia and Croatians, regardless of the place of publication and authorship,
- works in Croatian language,
- works published in Croatia.

Considering the fact that content of different quality and relevance is published on the internet, *Specific criteria* have been established related to the content, resource structure, reliability of the publisher, domain, format and uniqueness. *Content* criterion refers to a coherent, independent web content which has a permanent cultural, intellectual, scientific or artistic value Criterion of reliability and reputation of the *publisher or author* is also important. *Structure* criterion refers to the presentation of data such as title and publisher/issuing body responsible for content and creation of resources, design and arrangement of menus and data, as well as the regularity of updating. *Domain* criterion includes resources that are originally published on the .hr domain and are primarily selected for the web archive. In addition, resources on other domains (.com, .net, .info, .org etc.) may be selected if they meet other selection criteria (Fig. 2).

[4] Buzina, Tanja ; Willer, Mirna. Croatian digital Web archive: from project to service of the National and Univeristy Library in Zagreb. // Digitalia: rivista del digitale nei beni culturali. Available at: http://digitalia.sbn.it/article/view/437/277

[5] Content providers refers to producers and publishers

[6] Brautović, Marko. Razvoj hrvatskog online novinarstva 1993-2010. // MEDIANALI – znanstveni časopis za medije, novinarstvo, masovno komuniciranje, odnose s javnostima i kulturu društva 4, 8(2010), p. 29.

[7] Selection criteria // Croatian Web Archive. Available at: http://haw.nsk.hr/en/selection-criteria

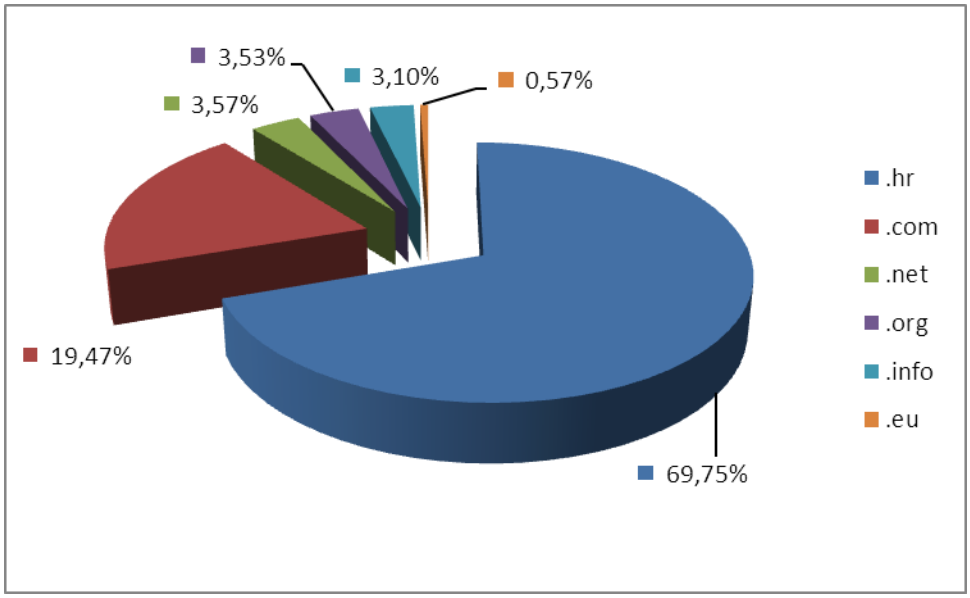

Fig. 2. Distribution of domains in HAW

Format criterion refers to the format that can be stored in the original form in order to keep the integrity and authenticity of the resource. Standard formats have priority. *Uniqueness* criterion gives priority to digital born resources.

Types of web resources stored in the Croatian Web Archive are integrating resources (websites of institutions, associations, clubs, research projects, news media, portal, blogs, official websites of counties, cities, etc.), serials (like for instance journals) and monographs (books). According to the last analysis of the content integrating resources account for the biggest percentage of 70%, followed by serials 10% and monographs 20 %.

The paper analyses and describes managing of 3581 titles of integrating resources. Although the workflow is the same for other types of resources in HAW like monographs and serials, the analysis shows diversity of integrating resources in HAW (Fig. 3)

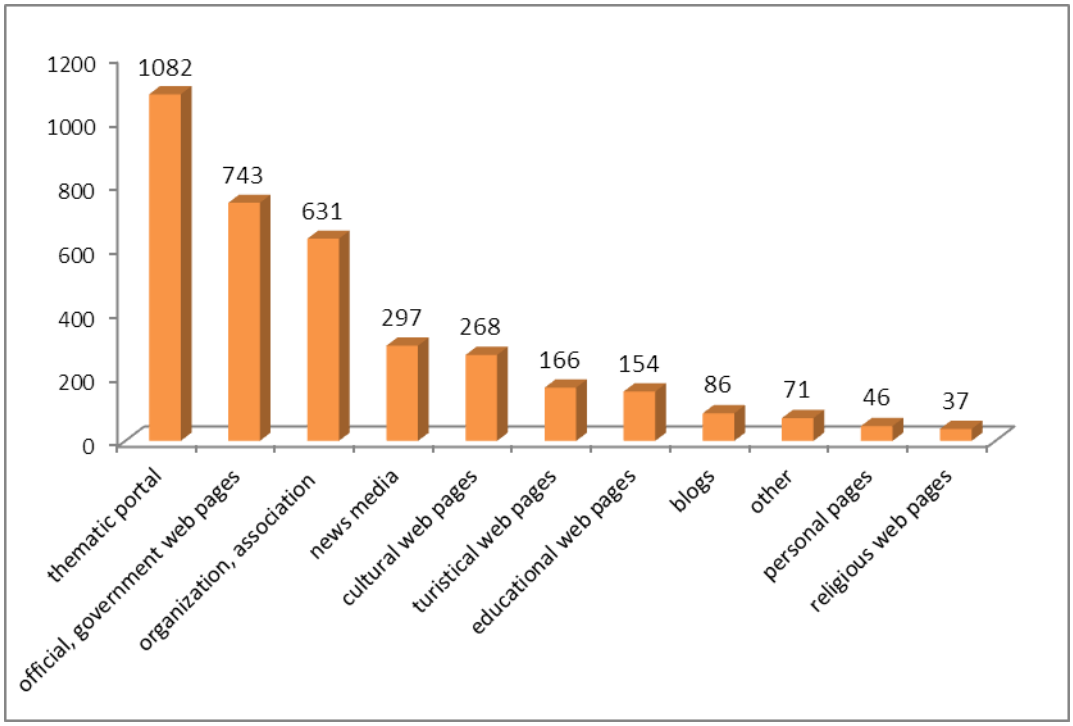

Fig. 3. Types of web resources in selective archive

From identification to archived copy (workflow)

There are several methods of identification of resources in HAW. The most common method is when librarians search and browse the web to identify a resource for cataloguing and archiving. Another one is application of online resources by website owners and/or content providers who have to fill the *Registration form*.[8] The *Registration form* is available on the HAW's homepage and contains information such as URL, title, statements of responsibility, access rights, and contact details. Once the application is received, library staff informs website owners and/or content providers whether their online resource is suitable for cataloguing and archiving according to the *Selection criteria*. Similarly, one of the methods of identification is a notification from the ISSN office for Croatia[9] which assigns ISSN to online integrating resources. The ISSN is assigned to news portals and thematic portals if they meet the additional criteria in the ISSN Manual i.e. present editorial content, and contain all the necessary data for description and identification etc.[10]

Continued cooperation of the HAW, website owners and/or content providers and ISSN Office resulted in 324 (almost 10%) archived websites with an ISSN, but also increased the number of registrations via *Registration form*, year after year. From 2010 to 2014 the annual number of registrations via the form has almost doubled (70 (2010), 136 (2014)) indicating increased awareness of website owners and/or content providers about the delivery of online legal deposit. Every website owner and/or content provider is notified of their website being harvested along with the information on the legal obligation of delivering the online resource and the name of the crawler for harvesting. Each website owner and/or content provider decides on the terms of access to their archived copy. Content archived in HAW can be publicly available through the HAW's website or if website owners and/or content providers deny public access, a minimum level of access is ensured within the NSK in a controlled working environment. By now just a few (0.14 %) website owners and/or content providers denied public access to the archived copy (Table 1). Each website owner and/or content provider can set HAW's logo on their website as a direct link to the archived copy of their resource in the Croatian Web Archive. The use of standardized web solutions contributes to the quality of harvesting. Therefore to ensure the better quality and ability to automatically create archived copies similar to the original, guidelines for creating, editing and technical recommendations that comply with the W3C recommendations were made for producers and/or owners.[11]

Bibliographic description of web resources

Once the title has been identified and selected according to the established criteria, the process of cataloguing begins. A manual for cataloguing integrating resources in MARC 21 format was created in line with the national cataloguing rules, as well as the current ISBD

[8] Obrazac za prijavu online publikacija. Available at: http://haw.nsk.hr/obrazac_za_prijavu_mreznih_publikacija

[9] Hrvatski ured za ISSN. Available at: http://www.nsk.hr/issn/#9

[10] ISSN Manual. Available at: http://www.issn.org/understanding-the-issn/assignment-rules/issn-manual/

[11] Preporuke za izradu mrežnih publikacija. Available at: http://haw.nsk.hr/preporuke_za_izradu_mreznih_publikacija

standards for continuing and electronic resources.[12] Because anybody can submit anything anywhere[13] and become an author and/or publisher, cataloguing and describing this type of resource is very challenging. Due to the dynamic, variable and unstable content is demanding and results in constant intervention in the bibliographic record. [14] The whole resource is taken as the base for description, but the current iteration is the source of information for description. Authority records are made for corporate bodies (web pages of organizations, institutions, associations and etc.), authors of blogs and personal pages and they are available in the Virtual International Authority File (VIAF). The new ISO number ISNI is assigned to some corporate bodies and authors.[15] A UDC is assigned to each title. UDC is used for subject categories displayed on the HAW's website.

The workflow is based on the interaction between ILS and the archiving system. Each catalogued resource that enters the archiving system is using predefined metadata. The process of exchanging metadata takes place at the same time on a daily basis; records from ILS enter the archiving system and the next day where the archiving process starts.

Software support for HAW is developed by Srce and is based on open source environment including: Debian Linux, MySQL, Oracle Java, Apache Tomcat, Apache HTTP Server and PHP. The goal was to build a system that would be able to gather and archive the selected web resource located within the Croatian web space, as well as preserving the original websites as much as possible.

In order to obtain the copy of the highest quality i.e. a copy as similar to the original as possible, archiving is assessed individually - a web curator (librarian) manually sets parameters, which ultimately results in a higher quality of archived copies.[16] The quality of archived content is obtained by adjusting the parameters that are responsible for the amount and quality of the content that needs to be archived. Gathering parameters for each resource are defined by a web curator. Each copy is manually checked and, if necessary, new parameters are added or depth of harvesting is changed, depending on the results of the first archived copy. Table 1 shows data on depth and common parameters of harvesting.

Some resources e.g., news media are archived daily. Frequency of harvesting is used to determine when resources will be queued for processing and when they can be set to be processed manually: days in a week, daily, day/s in a month and the months of the year (Table 1). The staff determines the frequency of harvesting of a certain resource in line with the importance of the resource for the general community, the importance of changes in content,, redesign, and the actual frequency of a resource.

[12] Buzina, Tanja; Holub, Karolina. Mrežna građa : upute za katalogizaciju u bibliografskom formatu MARC 21, izdanje 1999., 12. verzija dopuna. Zagreb: Nacionalna i sveučilišna knjižnica, 2011. Available at: http://haw.nsk.hr/upute_za_katalogizaciju

[13] A submission is not a publication.

[14] Holub, Karolina; Rudomino Ingeborg. Integrirajuća mrežna građa u Objedinjenom izdanju ISBD-a. // 16. seminar Arhivi, knjižnice i muzeji: mogućnosti suradnje u okruženju globalne informacijske infrastrukture: zbornik radova / uredili Nives Tomašević i Ivona Despot. Zagreb: Hrvatsko knjižničarsko društvo, 2013. Str.190

[15] Getliher, Danijela; Knežević-Cerovski, Ana. Međunarodni standardni broj imena (ISNI) u kontekstu normativnog nadzora // Vjesnik bibliotekara Hrvatske 57,1/3(2014), str. 37

[16] Due to constant change in web technologies we constantly improve basic components of the system, especially the gatherer with the new parameters like Seconds sleep after request, use content disposition header, auth type, username, password, handle flash which additionally contributed to the quality of harvested content. The first list of parameters for selective archiving are shown in: The Architecture of DAMP: A System for Harvesting and Archiving Web Publications http://haw.nsk.hr/arhiva/vol1/1210/6340/widwisawn.cdlr.strath.ac.uk/Issues/Vol3/issue3_3_1.html

Table 1. HAW's selective archive data

ISSN	Access	Depth of harvesting	Parameters of harvesting	Frequency of harvesting
With **9.09 %** Without **90.91 %**	Public **99.86 %** Restricted **0.14 %**	Depth 2 → **61.18%** Depth 3→ **26.96 %** Depth 1 → **6.1%** Other (0,4,5,6,7,8,)→ **5.76%**	Mandatory→ ***r****ecursion depth* Most common → *unwanted path pattern, always get embedded resources, alternative host, synonym* etc.	Daily (**0.08**%) Days in a month (**1.26**%) Days in a week (**0.17** %) Months in a year (**18.94** %) Manually (**79.55** %)

Selective approach to web archiving provides full control over management of web resources, but a labour intensive workflow shows a relatively small number of processed web resources in 10 years, over 5500 web resources.[17]

Full bibliographic description is the basis for the creation of metadata of archived titles to the NSK's metadata aggregator the European Library (TEL) and the Europeana. The HAW's metadata is delivered in MARC 21 xml format. HAW is by now the only web archive with its metadata available in the Europeana.

National (.hr) domain harvesting

Since the beginning of 1993 the national domain in Croatia has been managed by CARNet. In the period from 1992 to 2002 a total of 27,540 .hr domains have been assigned.[18] Unfortunately, in that period much of the web content on political, social and cultural contemporary life was lost.

Although the Croatian web space has been measured and monitored since 2002 technical, financial and organizational capabilities allowed the NSK to begin harvesting the national domain only in 2011. [19] In order to enlarge and improve the national collection of archived resources, NSK, in collaboration with its partner Srce, started harvesting the national domain (.hr) The harvesting was conducted with an open source crawler Heritrix developed by the Internet Archive. The workflow of domain harvesting differs from the selective archiving because the content is not bibliographically described and there is no interaction between ILS and archiving system. The process began with a seed list of URLs of all active .hr domains provided by CARNet. Due to the lack of disk space before any harvesting adjustment is required: the depth is limited to 4, the maximum number of resources per host is 50,000, the maximum file size is 100 MB, the depth of embedded resources is 3, respecting robot.txt rules[20]. The first harvesting has been conducted in the summer of 2011 when the initial seed

[17] An analysis of the cost of selective archiving approach has been made in 2008. Results are shown in Willer, Mirna; Buzina, Tanja; Holub, Karolina; Zajec, Jasenka; Milinović, Miroslav; Topolščak, Nebojša. Selective archiving of web resources: a study of processing costs. // Program: electronic library and information systems. 4,42(2008),341-364.

[18] Klarin, Sofija. Predmet, motivi i metode arhiviranja sadržaja weba // 8. seminar Arhivi, knjižnice, muzeji: mogućnosti suradnje u okruženju globalne informacijske infrastrukture: zbornik radova / uredila Tinka Katić. Zagreb: Hrvatsko knjižničarsko društvo, 2005. Str. 27

[19] Milinović, Miroslav. O mjerenju hrvatskog prostora weba // Vjesnik bibliotekara Hrvatske 48,2(2005), str. 27

[20] The robot harvests all content which allows harvesting to any robot, i.e. does not harvest content which has a restriction in robots.txt.

list contained 85,672 domains amounting to 4.1 TB. In period from 2011 to 2014 more than 17 TB of content have been harvested and data was stored in WARC file format with 266,240,439 resources with status code 200 OK.[21] Table 2 shows data for all four domain harvestings.

Table 2. Data from national domain harvestings

Year	Size (TB)	Number of .hr	Resources with a status code 200	Formats
2011	3,1	85,764	56,693,382	text/html, image/jpeg, image/png, image/gif, text/xml, application/pdf etc.
2012	4,1	74,812	60,903,245	
2013	4,6	76,944	69,123,642	
2014	5,7	85,721	79,520,170	

Thematic harvesting

From 2011 The Croatian Web Archive periodically harvests websites related to topics and events of national importance. Thematic harvestings include most common topics and themes like important national and political events, occurrences whose materials tends to disappear quickly from the web as well as unexpected situations in the world of politics, natural disasters etc. These initial lists of web pages were manually selected by HAW staff but public nominations are also taken into consideration where articles, parts and complete web pages, blogs, personal pages are selected. This content is not bibliographically described but each thematic collection consists of several metadata: title, size, extent (number of seeds/URLs) and description. Thematic harvesting is also carried out with crawler Heritrix and up until today five thematic harvestings have been conducted.[22]

What is good enough? (quality assurance)

An evaluation of a harvested website is an essential step in the web harvesting process. Over the years several modules for system monitoring have been developed as well as additional tools to help system librarians. *Checking large archived copies* is a tool that signals staff which archived copy exceeds size of 500 MB. *Daily report for possible duplicates* notifies of similarly collected samples. When the last two copies are similar in more than 80% it is likely that resources are online but not updated. It helps the staff to control older resources which disappeared or are not updated and are excluded from the automatic schedule. *Automatic monthly report* is a tool for checking the availability of the resource at its live URL. Fig. 4 shows percentage of active and/or disappeared web resources.

[21] Harvesting of the national web domain (.hr). Available at: http://haw.nsk.hr/en/harvesting-of-the-national-web-domain

[22] Thematic harvestings. Available at: http://haw.nsk.hr/en/thematic-harvestings

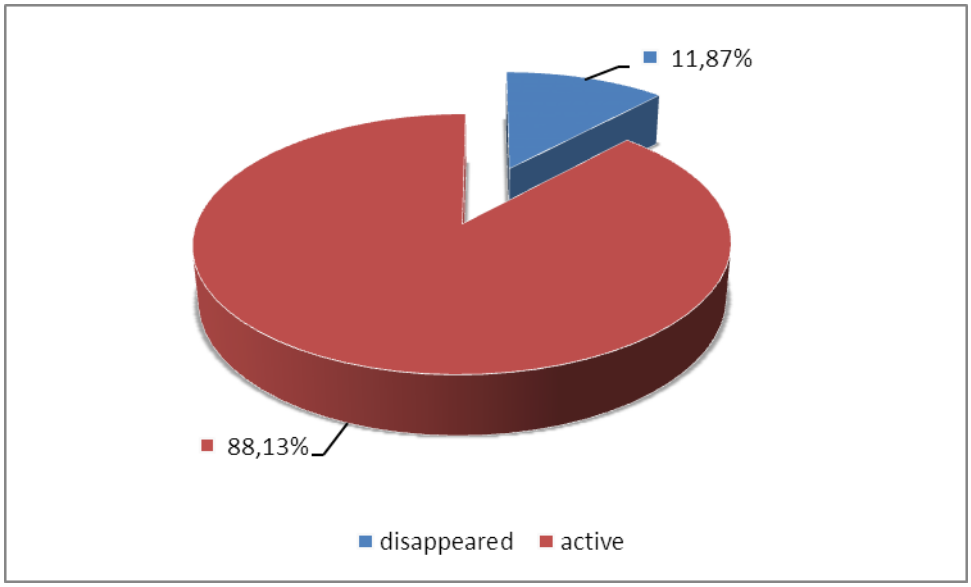

Fig. 4. Total percentage of active/ disappeared in HAW (2004-2015)

These technical tools provide good quality assurance in the archived copy of a website. To move forwards in developing software tools for selective harvesting an effort have to be made to archive websites (in Flash, JavaScript or Drupal) that are causing problems for HAW's crawler at the moment. Videos published on YouTube and similar streaming audio or video files are very difficult to harvest. In addition, some resources have not been archived (even though they meet the Selection criteria) due to their site design, site size or access restrictions imposed to HAW's robot.

In many cases the challenge is not only to harvest content but to archive it in a way that users can replay it and get the "look and feel" of the original content. New content type and web technologies demand adjusting and the crawler is continuously being changed and adjusted. A great deal of upgrading (up until now, 55 times) has been done since 2004 in order to keep abreast of the technological developments in this field.

In domain harvesting, quality assurance is more complicated and in a lower quality simply because the content is massive. Problems are causing poor configured websites and during harvesting wrongly generated links are deleted and added rules for better harvesting. In addition, at each domain harvesting the staff manually check a sample of 100 archived websites.

Access methods and usage in HAW

Since HAW collects publicly available content from the web, all archived content is publicly available and can be searched and browsed in several ways:

Selective harvesting

In HAW's selective part 99.86% of harvested content is publicly available (Table 1). Full-text indexing has been implemented by using the components from Java, Tomcat and Lucene. Search has been enabled and may be performed by any word in the title, URL, keywords. Advanced searching is also possible. In addition, users may browse HAW through subject categories that are extracted from UDC and alphabetically. Moreover, as each title has full bibliographic description they can be found through NUL's WebPAC as well.

Domain harvesting

The content collected in the past four .hr harvestings is publicly available via HAW's website. Archived copies are available through the Wayback Machine with interface in Croatian language.[23] Harvested websites can be accessed by entering the correct URL and choosing the year and date on the calendar in view.

Thematic harvesting

All thematic collections are publicly available on the HAW's website and can be searched through a title of a particular thematic collection to get the list of archived websites.

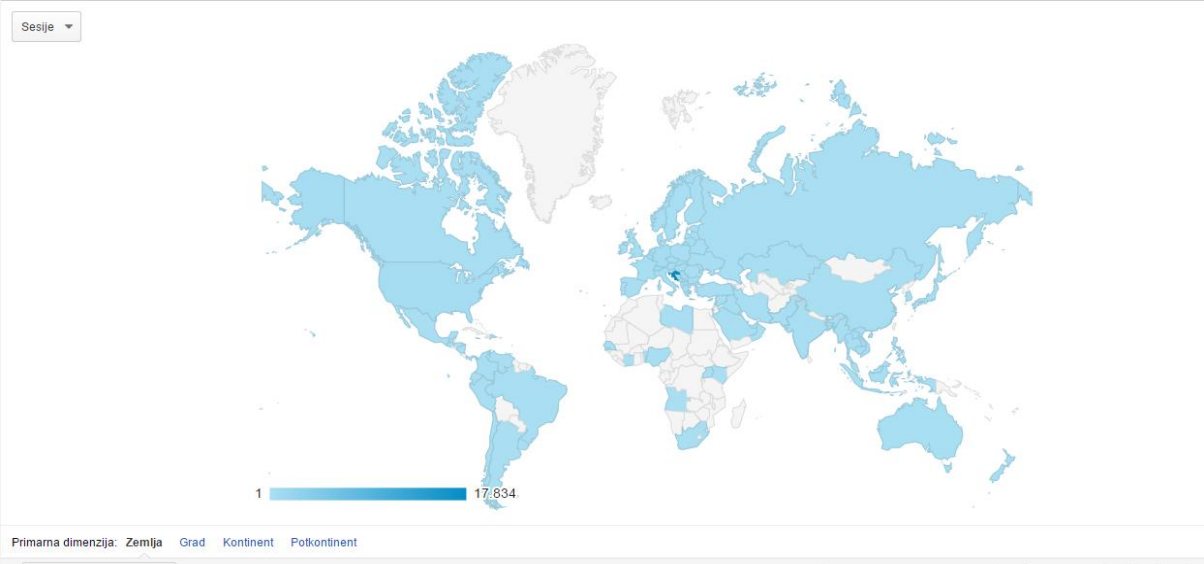

Fig. 5. Google Analytics access to the HAW by country in 2014

Usage

The staff monitors the use and behaviour of the users via Google Analytics, AWStats and Webalizer. Fig. 5 shows data (locations) from 2014 where HAW visited 21,783 users from all over the world, who stayed an average of 1:53 minutes and viewed more than 46,980 pages. [24] The most viewed pages are: *Naslovnica, O arhivu weba, Obvezni primjerak, Slobodna Dalmacija, Obrazac za prijavu online publikacija etc.*

Future plans

The Croatian Web Archive is a publicly available service with more than 25 TB of archived content

Although the National and University Library in Zagreb has been using diverse approaches to harvesting contemporary part of the cultural and scientific heritage for future generations for ten years, a certain amount of content is still irretrievably lost. A significant part of the public life of one country today takes place on various networking platforms i. e. social networks that accumulate different types of content and media, and are often used for business, research and communication. In that context it is a priority to broaden the Croatian web harvesting program with harvesting social media which have become globally important today as well for tomorrow's public and scholar research and usage.

[23] Wayback Machine interface. Available at: http://haw.nsk.hr/wayback/
[24] Source: Google Analytics

Acknowledgments

We would like to thank our colleague Draženko Celjak from University of Zagreb University Computing Centre (Srce) who supported us in data analysis.

References

Buzina, Tanja ; Willer, Mirna. Croatian digital Web archive: from project to service of the National and Univeristy Library in Zagreb. // Digitalia: rivista del digitale nei beni culturali. Available at: http://digitalia.sbn.it/article/view/437/277

Buzina, Tanja; Holub, Karolina. Mrežna građa : upute za katalogizaciju u bibliografskom formatu MARC 21, izdanje 1999., 12. verzija dopuna. Zagreb: Nacionalna i sveučilišna knjižnica, 2011. Available at: http://haw.nsk.hr/upute_za_katalogizaciju

Brautović, Marko. Razvoj hrvatskog online novinarstva 1993-2010. // MEDIANALI – znanstveni časopis za medije, novinarstvo, masovno komuniciranje, odnose s javnostima i kulturu društva 4, 8(2010) p. 24-42

Holub, Karolina ; Rudomino, Ingeborg. Croatian Web Archive : an overview. // Review of the National Center for Digitization, br. 25(2014), 11-16. Available at: http://elib.mi.sanu.ac.rs/files/journals/ncd/25/ncd25011.pdf

Hrvatski arhiv weba (HAW). // Nacionalna i sveučilišna knjižnica u Zagrebu. Available at: http://haw.nsk.hr

Willer, Mirna; Milinović, Miroslav. Prema trećoj generaciji knjižnično-informacijskih sustava : hibridna knjižnica za hibridne usluge // 8. seminar Arhivi, knjižnice, muzeji : mogućnosti suradnje u okruženju globalne informacijske infrastrukture : zbornik radova. Zagreb : Hrvatsko knjižničarsko društvo, 2005., str. 47.

Getliher, Danijela; Knežević-Cerovski, Ana. Međunarodni standardni broj imena (ISNI) u kontekstu normativnog nadzora // Vjesnik bibliotekara Hrvatske 57,1/3(2014), str. 37-50

Klarin, Sofija. Predmet, motivi i metode arhiviranja sadržaja weba // 8. seminar Arhivi, knjižnice, muzeji: mogućnosti suradnje u okruženju globalne informacijske infrastrukture: zbornik radova / uredila Tinka Katić. Zagreb: Hrvatsko knjižničarsko društvo, 2005. Str. 22-35.

Milinović, Miroslav. O mjerenju hrvatskog prostora weba // Vjesnik bibliotekara Hrvatske 48,2(2005), str. 26-34

Selection criteria // Croatian Web Archive. Available at: http://haw.nsk.hr/en/selection-criteria

Willer, Mirna; Buzina, Tanja; Holub, Karolina; Zajec, Jasenka; Milinović, Miroslav; Topolščak, Nebojša. Selective archiving of web resources: a study of processing costs. // Program: electronic library and information systems. 4,42(2008),341-364.

Zakon o knjižnicama. (1997) Avilable at:
http://narodne-novine.nn.hr/clanci/sluzbeni/267274.html