Преглед НЦД 25 (2014), 11–16

Karolina Holub, Ingeborg Rudomino
National and University Library in Zagreb, Croatia

CROATIAN WEB ARCHIVE: AN OVERVIEW

Abstract: The National and University Library in Zagreb (Library) established the Croatian Web Archive (HAW) in 2004 in collaboration with the University of Zagreb Computing Centre (Srce). The objective of the HAW is to collect, preserve and provide access to web resources that are considered an important part of Croatian cultural heritage. This paper presents the Library's practice in archiving Croatian web resources. Croatian Web Archive started with a selective approach of collecting web resources. In order to enlarge and improve the national collection of archived resources, in 2011 the National and University Library in Zagreb decided to undertake the whole national domain harvesting. In addition, thematic harvestings focusing on particular subjects or events (e.g. national elections, sport events or natural disasters, etc.) were conducted as part of the HAW's activities.

Keywords: Croatian Web Archive (Hrvatski arhiv weba - HAW), web archiving, selective archiving, national .hr domain harvesting, thematic harvesting

1. Introduction

The web plays an important role in society, it changes the way of living, learning, communicating constantly and permanently. Content published on the Internet represents a significant part of contemporary cultural and scientific heritage. Development of Internet and World Wide Web confronts libraries with a challenge to identify, catalogue, preserve and enable access to this type of content. Many diverse resources that memory institutions collect, such as scholarly resources, campaign materials, works of art, government documents, books, journals, news etc. are now available only on the web. Due to the dynamic nature of the web this content is constantly changing and a great number of websites have a quite short lifespan of approximately 44 or 75 days [2]. Some of these resources have already disappeared forever and web archives devise various ways of capturing the web and keeping copies as evidence of a time.

To make sure this content survives for the next generations it has to be captured at the time it is published. In 1996 a non-profit organization and digital library Internet Archive with several national libraries began archiving the web. They established their own web archives, developed tools, monitored trends, encouraged and provided support worldwide in web archiving and preservation. The general concept was to collect a sample of world's online society and preserve it for future generations. Among the first national libraries that started archiving the web were the Library of Congress, the National Library of Australia, the National Library of Sweden, the National Library of Norway and the National Library of New Zealand.

The National and University Library in Zagreb is a memory institution responsible for collecting, bibliographic description, storage and providing access to all types of resources published in Croatia. Early recognition of the significance of collecting and storing online content made web archiving one of the core Library's activities in 2004.

2. About HAW

The National and University Library in Zagreb in collaboration with the University of Zagreb Computing Centre (Srce) established in 2004 a system for collecting and archiving the legal deposit copy of Croatian web resources. The aim was to preserve, to the largest measure possible, the original contents, formats and functionalities. The entire program was based on the fact that in 1997 Croatia passed a law by which *online publications* are included within legal deposit and each publisher is legally bound to deliver to the Library a legal deposit copy of a resource published online [8].

The system for collecting and managing web resources was named Digital Archive of Croatian Web Publications (DAMP). From 2004 to the present, the archiving system was developed in line with the development and changes in web technologies.

In 2010 the National and University Library in Zagreb changed the name of the archive[1] to HAW (*Hrvatski arhiv weba*, Croatian Web Archive). The new name describes the purpose of archiving web resources more precisely and does not mislead users about the scope of the service. Also, the word "digital" in the title was redundant because the web implies digital form. Moreover, the new name is compliant with similar names of other archives, such as UK Web Archive, Finnish Web Archive and Iceland Web Archive.

Up to 2010 the content of the HAW was integrated in the Library's website. Based on research, surveys and practices it was concluded that users had difficulties in finding information about the HAW and often mistook it for digitized materials. At the end of 2010 the new HAW website was finished with a lot of information, documents and a new range of functions [4]. The new website enabled full text search by title, URL and keywords, browse by subject areas and browse by alphabetically sorted titles. In addition, new search terms in the form of tag cloud and a list of the last five archived resources have been added (Figure 1).

Figure 1: Homepage of the Croatian Web Archive

In order to broaden the scope of the national collection of archived resources, in 2011 the HAW conducted the national (.hr) domain harvesting for the first time, and began with thematic harvesting of web content related to important national events.

[1] Although the name DAMP was kept for internal purposes.

From 2012 metadata from the HAW can be found through the largest European digital library Europeana [3]. The HAW is for now the only web archive whose metadata are a part of Europeana.

Today the HAW primarily collects digital born content (such as important sport competitions, political, cultural and other popular events) with particular care, as it documents everyday life.

3. Selective archiving

The National and University Library in Zagreb has been archiving web resources selectively since 2004. Each web resource is assessed, bibliographically described and archived according to the pre-established selection criteria (general and specific). General criteria apply to works by Croatian authors published in Croatia and abroad, works about Croatia and Croatians, works in Croatian and works published in Croatia. Specific criteria refer to the content, publication structure, reputation and reliability of the publisher, domain, format and uniqueness of the web publication considered for cataloguing and archiving.

The workflow starts with search for new resources, continues with selection, followed by bibliographic description and archiving process (Figure 2).

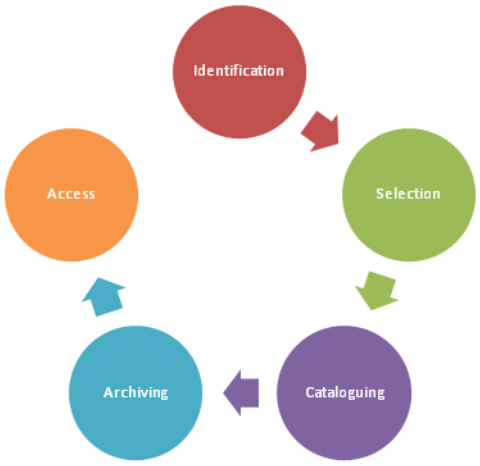

Figure 2: Workflow of the selective archiving

Web resources that are possible candidates for the archive are mostly identified and selected by the Library staff using search engines (e.g. Google). Information about new web resources also comes from other Library units such as ISSN Center for Croatia. Publishers and authors also notify the HAW about the existence of a web resource by filling out the Registration form at the HAW's web site.

The functionality of the web resources management system is based on the process of interaction between the Library system and the DAMP archiving system. Predetermined resource metadata, previously bibliographically described, are automatically transferred from the Library system to DAMP system where the archiving process starts.

The software for selective archiving was developed by the University of Zagreb Computing Centre (Srce) and named DAMP software. It is based on open source development and production environment, operation system used is Debian Linux 2.6 with the MySQL 5.0, Java 1.6, Apache, Tomcat 6.0, Apache/2.2 and PHP 5.2 environment.

After successful archiving each publisher is notified about bibliographic description and archiving. Online access to archived resources is free and if public access is denied by the publisher, minimum level of access to the resource is available, i.e. one authorized user within the Library in a controlled working environment.

Types of resources contained in the HAW are integrating resources (e.g. news portals, web pages of institutions, associations, blogs etc.), monographic publications (e.g. books) and serials (e.g. journals).

The HAW is fully integrated with the Library information system and every archived resource is publicly available via OPAC and the HAW's website http://haw.nsk.hr/ [7].

In November 2013 the size of the Archive was 6.1 TB, 33,971 archived instances, more than 144,385,915 files and 5,010 unique titles.

4. National domain and thematic harvestings

4.1. National .hr domain harvesting. Taking into account that the Library has the responsibility to preserve Croatian social, scientific and cultural history, the importance of taking a snapshot of all publicity available resources under the national top level domain (.hr) has been recognized. The purpose of domain crawl is to collect, store and provide public access to online Croatian cultural and scientific heritage.

The first domain harvesting started in July and August 2011 and was performed by the National and University Library in Zagreb and the University of Zagreb Computing Centre (Srce). The crawl begun with a seed list of URLs of all active .hr domains provided by Croatian DNS Service at the Croatian Academic and Researchers Network (CARNet) [6]. The initial seed list contained 85,672 URLs. A robot was developed for checking if an active website really existed behind the domain name. The final seed list created for harvesting contained 62,459 URLs [1]. The harvesting was conducted with open source crawler Heritrix developed by Internet Archive [5]. The main reason for this decision was that Heritrix allows modifying a large number of settings, including crawling priorities, filters, robot politeness and ability to store the resources in WARC format. Between 18th July and 18th August over 56,000,000 files were collected resulting in 3.1 TB of data.

The second domain harvesting was carried out from 19th until 31st December 2012. The initial seed list contained 74,812 URLs, but after the robot check the active seed list amounted to 60,639 URLs. This time 60,903,245 files were collected resulting in 4.1 TB of compressed data.

Before the beginning of each harvesting, the Internet community (webmasters) was notified through the Library and HAW web sites and social networks with the purpose of harvesting as well as with the name and the settings of the robot.

So far the harvestings were generally successful and improved the public awareness about the HAW. The online heritage will be a valuable resource for future researchers, historians, scholars as well for the entire community.

4.2. Thematic harvestings. Along with the annual harvestings of the whole national domain (.hr) and selective archiving, the HAW also decided to run thematic harvestings. These harvestings are conducted throughout the year and often concentrate on specific subjects or timely events. The most common topics and themes of the harvests are, for example, important national and political events, whose materials tend to disappear quickly from the web after the event is finished, as well as unexpected situations in world of politics, natural disasters etc. The purpose of thematic/event harvestings is to anticipate future research needs and supplement any areas missed in the national domain and selective harvestings.

The initial lists of web pages were manually selected by the HAW staff and also carried out with crawler Heritrix. The process starts with web browsing to discover pages or whole web sites related to the selected topic. Public nominations are also taken into consideration but usually librarians are those who recommend the seeds.

By the end of the 2013 five thematic harvestings were carried out (Figure 3). In 2011 the first thematic collection was on the result of the *2011 Croatian Parliament Elections*,

harvesting conducted from 2nd till 8th December 2011. 140 URLs in total size of 64 GB were collected. The second collection pertained to the *Referendum on the Accession of the Republic of Croatia to the European Union (22nd January 2012)*. The harvesting was conducted on 16th October 2012. 476 URLs in total size of 929 MB were collected.

The third thematic collection was *2013 European Parliament Elections in Croatia*. The harvesting was conducted on 1st May 2013. 55 URLs in total size of 8.9 GB were collected.

The fourth thematic collection was dedicated to *2013 Local Elections* and included regular elections of members of representative bodies of local and territorial (regional) government. The harvesting was conducted from 6th to 7th June 2013 and 202 URLs were collected with total size of 15 GB.

The most recent and probably one of the most important collections is *The Accession of the Republic of Croatia to the European Union (1st July 2013).* The harvesting was conducted from 16th until 17th July 2013 during which 116 URLs were collected in total size of 32 GB.

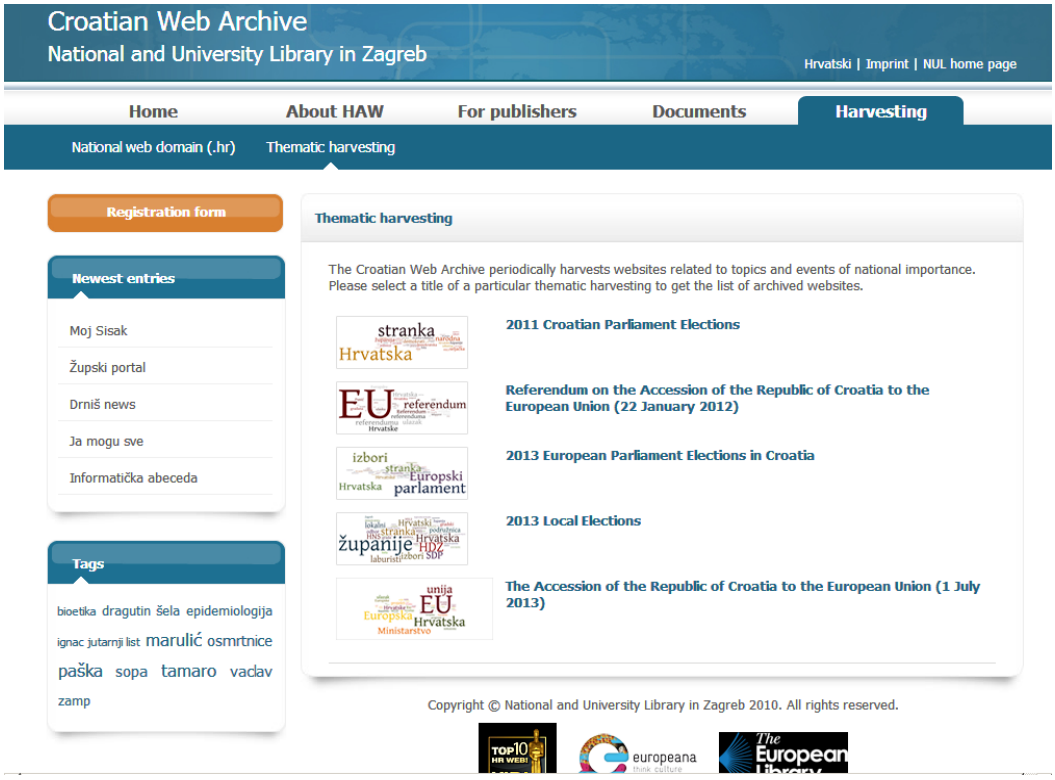

Figure 3: Screenshot of the conducted thematic harvestings

From 2013 all domain and thematic crawls are available to the public via the HAW's website http://haw.nsk.hr/en/harvesting-of-the-national-web-domain.

5. Future plans

Full potential of web archives are yet to be exploited. By now the National and University Library in Zagreb has established a sustainable, functional and publicly accessible web archiving system with more than 13 TB of archived content. The most important goal is to continue with present work of selective archiving of significant Croatian websites, annual harvestings of the whole national domain and keep running thematic/event harvestings when required.

In addition, the HAW would like to extend harvestings outside national .hr domain such as .org, .net, .com, .info, etc.

Until recently, every archiving institution was focused just on data collection and not on usage. Considering that the content of the HAW is publicly available, our plan is to find a way of bringing the HAW closer to the researchers and scholars by evaluating the current and developing new tools and methods for increasing the usage of the HAW.

Diverse institutions and countries are joining their forces and working together in building collaborative collections throughout the world. Knowing that institutions and libraries in the region are not involved in web archiving we started to think about regional collaboration in the common interest.

References

[1] Celjak, D., Milinović, M. *Harvestiranje hrvatskoga weba: arhitektura programskoga sustava za harvestiranje i iskustva stečena njegovom upotrebom*. 15. seminar Arhivi, knjižnice i muzeji: mogućnosti suradnje u okruženju globalne informacijske infrastrukture: zbornik radova. Zagreb, Hrvatsko knjižničarsko društvo, 2012. Pp. 144-160.
[2] Day, M. *Collecting and preserving the World Wide Web*. Available at: http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf (2013-10-15)
[3] *Europeana*. Available at: http://www.europeana.eu/ (2013-10-29)
[4] *Hrvatski arhiv weba*. Available at: http://haw.nsk.hr/ (2013-10-15)
[5] *Internet Archive*. Available at: https://archive.org/ (2013-10-29)
[6] *Registar .hr domena*. Available at: http://www.dns.hr/ (2013-10-29)
[7] Willer, M., Buzina, T., Holub, K., Zajec, J., Milinović, M., Topolščak, N. *Selective archiving of web resources: a study of processing costs*. Program: electronic library and information systems 4, 42(2008), 341-364.
[8] *Zakon o knjižnicama*. Available at: http://narodne-novine.nn.hr/clanci/sluzbeni/267274.html (2013-10-15)

kholub@nsk.hr
irudomino@nsk.hr