Web museum, web library, web archive

The responsibility of public collections to preserve digital culture

Márton Németh*†, László Drótos*

* National Széchényi Library. H-1014, Szent György tér 4-5-6, Budapest, Hungary
† University of Debrecen Doctoral School of Infomatics. H-4028, Kassai út 28, Debrecen, Hungary
National Széchényi Library. H-1014, Szent György tér 4-5-6, Budapest, Hungary
E-mail: {nemeth.marton drotos.laszlo}@oszk.hu
nmarton@mailbox.unideb.hu

Abstract: This paper introduces the international context of web archiving in general. It offers a short summary about the relevance of web archiving to libraries, archives and museums. It also provides a brief overview about the web archiving pilot project in the Hungarian National Széchényi Library. Basic conception and goals are being described in this context.

Keywords: World Wide Web; Web-Archiving; international web-archiving efforts; long-term digital preservation; web archiving pilot project; National Széchényi Library

Introduction

In Hungary, the acquisition of born-digital books started in 1994 and an e-journal collection established in 2003. However, web sites and web materials have not been the integral part of the national library's digital collection. Only some institutional tests were established in this field (Kampis & Gulyás, 2013). In May 2017 by the framework of the National Library System Development Programme a web archiving pilot project (National Széchényi Library, 2017) has started at the National Széchényi Library (NSZL). On this paper, we would like to offer an introduction into the field of web-archiving with its international context and describe briefly the Hungarian pilot project.

Why Web Archiving? Current challenges, efforts, international context

Archiving the world-wide web is essential to preserve the status of the online world for future generations. Without internet archiving it will not be possible to research the history of the virtual world, and the online appearance of real world events. Analysing and describing large digital datasets harvested from the web have become a major research field. Web archiving is also needed to reference virtual resources permanently in scientific publications and in education materials. Web archiving efforts covers the whole public collection sphere

including libraries, archives, museums with different kind of specific tasks. Not just the archiving of digital objects is essential but we have to preserve the software and hardware architectures in order to emulate them perhaps hundred centuries later on virtual machines. UNESCO Charta about the digital cultural heritage states that “The digital heritage consists of unique resources of human knowledge and expression. It embraces cultural, educational, scientific and administrative resources, as well as technical, legal, medical and other kinds of information created digitally, or converted into digital form from existing analogue resources. Where resources are ‘born digital’, there is no other format but the digital object. Many of these resources have lasting value and significance, and therefore constitute a heritage that should be protected and preserved for current and future generations. This ever-growing heritage may exist in any language, in any part of the world, and in any area of human knowledge or expression. The world’s digital heritage is at risk of being lost to posterity. Contributing factors include the rapid obsolescence of the hardware and software which brings it to life, uncertainties about resources, responsibility and methods for maintenance and preservation, and the lack of supportive legislation. Attitudinal change has fallen behind technological change. Digital evolution has been too rapid and costly for governments and institutions to develop timely and informed preservation strategies. The threat to the economic, social, intellectual and cultural potential of the heritage – the building blocks of the future – has not been fully grasped.” (UNESCO 2003). UNESCO advices to establish strategies, policy guidelines, training programmes in order to reach our goal.

The current situation is not really ideal. David Rosenthal has just written an article about the status of the current web archiving efforts. He is quite afraid that we can lose this battle. He points out that granting the necessary recourses is mainly an economic challenge. “With an unlimited budget collection and preservation isn’t a problem. The reason we’re collecting and preserving less than half the classic Web of quasi-static linked documents is that no-one has the money to do much better. The other half is more difficult and thus more expensive. Collecting and preserving the whole of the classic Web would need the current global Web archiving budget to be roughly tripled, perhaps an additional $50M/yr. Then there are the much higher costs involved in preserving the much more than half of the dynamic ‘Web 2.0’ we currently miss.” (Rosenthal, 2017).

Nowadays the International Internet Preservation Consortium (IIPC) is the organisation that offers a solid international framework for most of the stakeholders that are active in web archiving fields in order to make joint efforts to better policies, guidelines, training

programmes in this field. (IIPC, 2017). The other main international actor in this field is the Internet Archive (“Internet Archive,” 2017.) It is a private non-profit company that has already harvested Hungarian content of the world wide web. They are searching for collaborative partners, like national libraries to run the complex tasks of web archiving in a shared platform. Internet Archive is taking snapshots from the selected webpages. The problem is that these snapshots cannot be display individually by their Wayback Machine software, but a comprehensive overview is being offered based on different snapshots made in different periods. The question of authenticity is a central issue in this context. The libraries can pay more attention to this aspect of web archiving than a private company does. Currently around 40 national web archives exists in more than 30 countries. In some cases, the National Library is responsible for these tasks (for example in Austria, Czech Republic, Denmark, Sweden). In other countries a national consortium is responsible for web archiving efforts (Australia, United Kingdom). In some other cases a loose coordination exists among different cultural institutions that are active in this field including the National Library and the National Archive (for example in the Netherlands).

Web Archiving activities in the context of cultural heritage collections

As we briefly mentioned above web archiving activities can effectively fits to the portfolio of libraries, archives and museums in different ways.

The web archiving projects in libraries ideally set by the legal deposit law. The collection guidelines are focusing on the public content of the web as a kind of publication resource. Materials must be selected and enriched with metadata. The public access of these selected materials depends on the actual legal environment. Selection can be topic-based or event-based. In almost all countries with legal deposit law libraries can select web documents without restrictions however the permission of the content-owner needed to offer them in a public service environment. Without permission these materials can be accessed through the library intranet on dedicated terminals only for archivists and researchers. The selectively harvested materials can be an integral part of the national bibliography and the library catalogue. Offering permanent linking to various resources is also an essential goal. Some general harvest as a snapshot of the current status of a national segment of the web can also be done, however this collection cannot be used publicly due to the copyright law, it is mainly a preservation task to the future.

The web archive of the archives mainly consists non-public, unpublished materials (e-documentations of companies, institutions, intranet, non-public groups and websites on the internet). The other main resources are the publicly available personal materials (forums, personal blog, photos, videos, social network pages and channels etc.). These web materials can be managed to a set of digital fond perhaps together with offline digital materials. A special archive task is the collection and preservations of government and public administration based digital materials, websites and other online materials.

A web archive of a museum is the main repository of the web objects related to fine arts and industrial arts. The web presence of artists also can be archived. Museums can be builds up selectively harvested web collections that fit to the institutional profile (for example local history web collection local events broadcasted online, photo collection, museum of sports, e-commerce based materials, web documents related to different kind of technologies and tools etc.). Museums can be primary places to historical research based on web archive materials. The first examples through cases studies of concrete projects of these kind of research activities can be found (Brüger et al. 2017). This book offers a good introductory overview about web archiving and the case studies offers a really good inside of various collaboration forms with linked data research field and various disciplines in arts and humanities.

Internet Archiving Pilot project in the National Széchényi Library

Based on the international experiences our main aim is to build-up a workflow, that can function permanently from 2019 to archive the Hungarian segment of the web. The main aim of the pilot project is to test the available software tools, administration software products and available hardware. Software tools are all open source products. The main challenge is that only a limited number of people are developing open source software in this field with restricted capabilities. To build up our own portfolio a high level of customization of different software products is highly needed.

We do not want to collect audiovisual materials as it is out of scope of the library collection portfolio. Hopefully the Hungarian Audiovisual Archive will be active in this field.

A main issue by the establishment of a web archiving workflow is the setup of the harvesting policy. We would like to focus on three types of harvesting. General harvesting is offering a snapshot twice a year from the Hungarian segment of the web. It includes a representation of the content under the .hu domain and a set of websites in foreign servers related to Hungary

and to the Hungarian culture. Event-based harvesting is focusing the online appearance of special events through selected websites (for example parliamentary elections). By selective harvesting a curated set of websites are being harvested regularly that contains cultural, educational, scientific, social and political topics. Building partnerships with several online content providers and memory institutions is essential to select and curate the born-digital resources.

As the first phase of the pilot project, we are currently testing the capabilities of the Heritrix 3.3 and the Wayback-Machine software to harvest and display a small segment of Hungarian library homepages or other cultural websites and electronic journals. The pilot of general harvesting of the .hu domain is being planned from the end of 2017 till the spring of 2018.

We are planning to implement the Memento protocol to offer our harvested content for joint search and retrieval services to establish fruitful collaboration with foreign web archives.

We have established a temporary pilot project homepage with a wiki of the important factual data in web archiving field (National Széchényi Library, 2017). We are offering a large selection of international articles in a bibliography. The members of the pilot project groups are taking presentations, writing articles to professional journals, organized a workshop to focus on the broader context of web archiving and introduce the project to broader audience. Another main future task is to provide educational materials and organize trainings in web archiving field. a course has planned in cooperation with the Hungarian Library institute for cultural heritage professionals and hopefully will be introduced in 2018. Furthermore we are a member of the training group of the IIPC that coordinates the international efforts in web archiving education field.

Followed by the pilot period similarly to most of the national library practices we hope that a model for permanent web harvesting can be implemented from 2019.

Epilogue

The rapidly developing domain of web archiving makes a central importance in order to preserve our digital cultural heritage. We have already lost a significant part of our web heritage and it is our common responsibility to set the necessary conditions of long-term web preservation to prevent the emergence of a digital dark age from a future perspective.

References

Brügger. N., Schröder. R. (Ed.). (2017). *The web as a history*. London: UCL Press

Internet Archive. (n.d.). Retrieved October 9, 2017, from http://www.archive.org

Internet Preservation Consortium. (n.d.). Retrieved October 9, 2017, from http://www.netpreserve.org

Kampis, G., & Gulyás, L. (2013). Big is small, and changes slowly in Hungary. In Coginfo 2013 Conference.

National Széchényi Library. (2017). OSZK web archiving pilot project portal. Retrieved October 9, 2017, from http://mekosztaly.oszk.hu/mia

Rosenthal, D. (2017). Losing the battle to archive the web. Retrieved December 18, 2017, from http://dpconline.org/blog/idpd/losing-the-battle-to-archive-the-web

UNESCO Charter on the Preservation of Cultural Digital Heritage. Retrieved December 18, 2017, from http://portal.unesco.org/en/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html