The education of web archiving

László Drótos <drotos.laszlo@oszk.hu>
Márton Németh <nemeth.marton@oszk.hu>
(National Széchényi Library, Budapest, Hungary)

The education of web-archiving

The article is focusing on three main issues. At first, an overview is being offered about an online research seminar for PhD students and web-archiving professionals organized by the NETLAB Research group, Aarhus University, Denmark. Secondly, the recently established Education and Training Working Group of the IIPC consortium is being introduced. A quick overview is being offered about a brief survey on best web archiving education practices and future. Thirdly, a Hungarian web-archiving training concept is being described. The training will be organized by the Library Institute for any kind of cultural heritage professionals that want to get basic skills and competences in this field.

keywords: education, web archiving, e-learning, NETLAB, IIPC

Introduction

Our paper is covering three main topics. Firstly, we are offering a short overview about an online seminar that was organized for PhD students and professional experts about web archiving by the NetLab research group, Aarhus University, Denmark. Secondly, the initial activities of the newly established Training Working Group of International Internet Preservation Consortium (IIPC) being introduced. A comprehensive survey made by this working group focused on best practices, current experiences and plans for the future. The first results are being presented. Thirdly we are introducing the initial plan and curriculum of a web archiving course in collaboration with the Library Institute. The course will focus on public collection professionals offering them an overview about tools and methodology of web archiving.

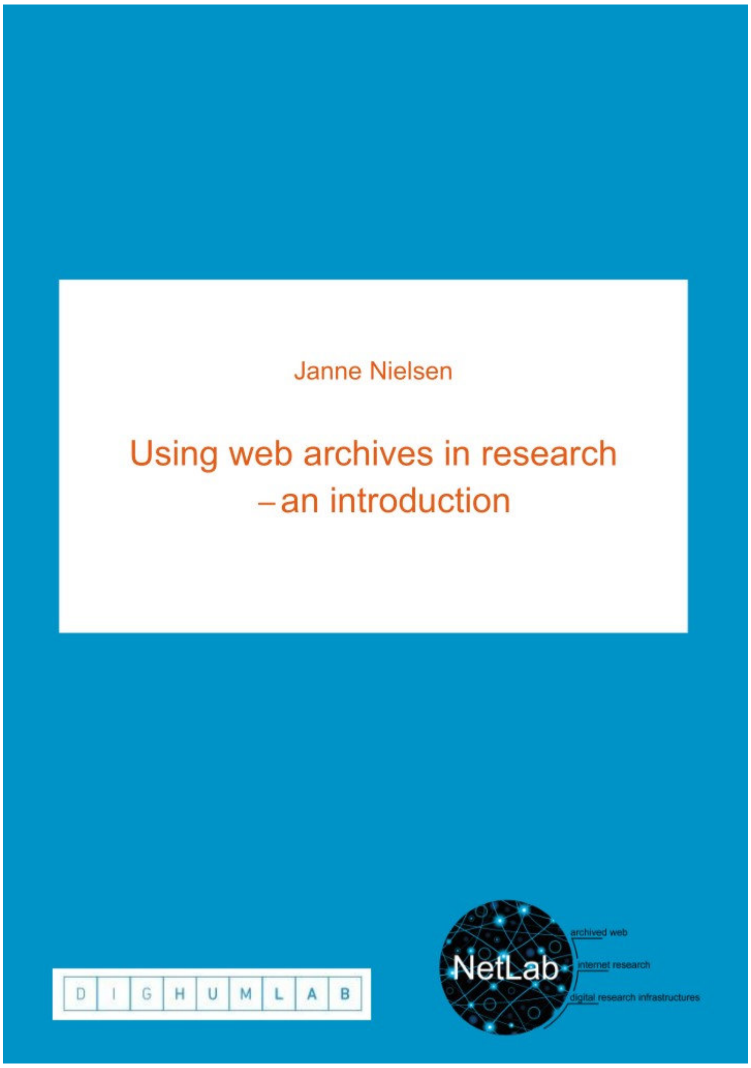
1. NetLab online course

The NetLab research group at Aarhus University, Denmark is a part of the national DIGHUMLAB[1] research infrastructure network led by prof. Niels Brügger. They have started to offer online courses about web archiving for two years. The major target groups are PhD students, public collection professionals and researchers. In the autumn of 2017 they introduced their first course entirely in English for an international audience [2]. The participation was free and it was really optimal to make new connections among the experts of the newly established web-archiving projects in Hungary and Belgium and work together with Danish colleagues as well. The online seminar was being held in a password protected

[1] More details about the educational activities of NETLAB konzortium, retrieved12.06.2018. http://www.netlab.dk/services/courses/

[2] Course description, retrieved 12.06.2018. http://www.netlab.dk/wp-content/uploads/2017/04/NetLab-Web-Archiving-Course-Brochure.pdf

Moodle-based e-learning interface. There we could access the exercises, training materials, hand-in the answers and make interactive conversations on the course forum. This interface was only available during the course but the participants could save all materials just after the finishing. The seminar covered five main topics A handbook by Janne Nielsen offered a general support through all topics and exercises..[3] *(Figure 1)*



Janne Nielsen

Using web archives in research
– an introduction

Figure 1 Cover of the coursebook

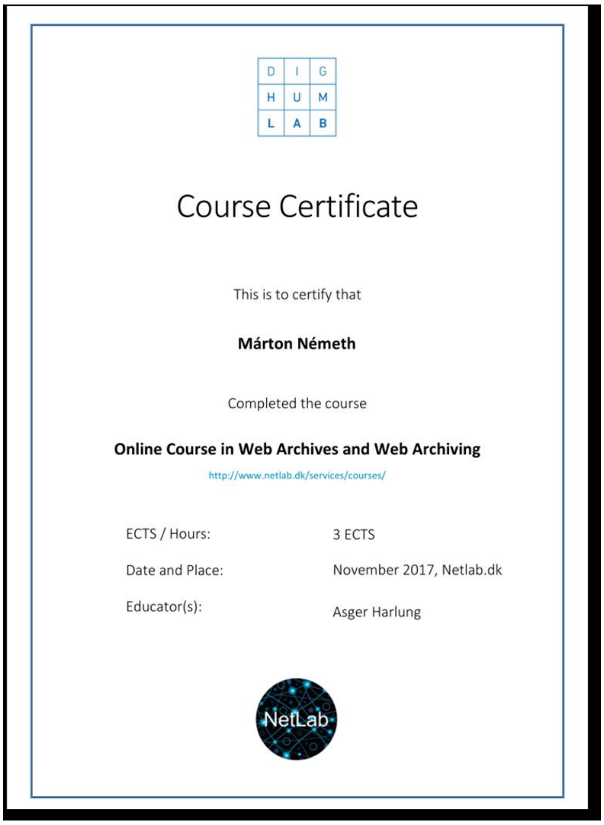
At the beginning of the course the professional background, expectations to the course and the types of involvement with web-archiving were discussed among the participants. The pedagogic style of the seminar is constructivist. It heavily relies on the professional profiles and level of involvement of the participants. In each semester the profile of the seminar by certain groups is largely different on this way. Followed by an overview about general interests in web archiving of the members, the second task was to specify professional topics, tasks and formulate a small-scale research plan related to web archiving. The third task was to find three websites by your own interests (the sites must be in operation for minimum one year) and describe the archiving challenges of them. Samples could be found in Internet Archive[4] or any publicly available archived site from a national web archive could be also

[3] Freely available , retrieved 12.06.2018. http://www.netlab.dk/wp-content/uploads/2016/10/Nielsen_Using_Web_Archives_in_Research.pdf

[4] Archive.org, hozzáférés 2018.06.12., http://www.archive.org

used. By the fourth task a collection strategy of websites had to be formulated. Followed by that the appropriate software background had to be selected and pilot harvests had to be initiated. Finally, the overall experiences had to be summarised in a report. This task appeared to be the most useful one because we could share experiences about pilot harvests with our Belgian colleagues and we could try to find answers about several archiving challenges. We could evaluate various software tools and analysing the harvest results. Our Danish colleagues from the Danish Web Archive[5] also could share with us their own experiences. The major goal was to formulate relevant questions about practical tasks in order to effectively formulate further of our own web-archiving pilot projects in Belgium and in Hungary. The final, fifth task was to make a general closing overview by discussions and filling up an evaluation survey made by the organizers. At the end all participants got an official certificate, demonstrating the completion of the course. *(Figure 2)*

In case of this web-archiving seminar the applied pedagogical style turned to be really effective. The theoretical background was available in written form for individual studying. The lessons based on this theoretical core were really practice-based focusing on specific tasks and challenges. A major aim was to ensure the long-term application of course experiences on our own job in an effective way. We could learn the most from each other. By planning tasks, discussing software problems, archiving issues and resolving some challenges made us really valuable experiences. The course effectively helped the foundation of our own web-archiving pilot project in the National Széchényi Library.



D I G
H U M
L A B
NetLab

Course Certificate

This is to certify that

Márton Németh

Completed the course

Online Course in Web Archives and Web Archiving

http://www.netlab.dk/services/courses/

ECTS / Hours:	3 ECTS
Date and Place:	November 2017, Netlab.dk
Educator(s):	Asger Harlung

Figure 2 Course Certificate

[5] Netarkivet, retrieved 12.06.2018., http://www.netarkivet.dk

2. IIPC Training Working Group (IIPC TWG)

Members of the IIPC international consortium are public and private organizations, institutions that are preserving online materials.[6]. Primary tasks of the consortium are the development of technologies, methodologies, standards related to web archiving, sharing national best practices, supporting international collaboration, granting the broad access to the archived web materials and helping to re-use these datasets in various ways. The Training Working Group (TWG) had established at the end of 2017.[7]. By their first project a survey was compiled[8]. The main aim was to collect basic information about national web archiving projects: Who, Where and in what kind of frameworks are working with web-archiving. The survey was also focused on human background of each institution and the aims and needs of professionals in education and training aspects of web archiving The survey was open in January, 2018. A quick summary of the results can be presented[9].

The answer of respondents to the survey was 224, representing a global professional group from five continents. Web archiving activities have mainly done by universities, research institutes and in a smaller but relevant scale: national libraries. The number of archives with web archiving activities is also relevant furthermore we can find museums, audio-visual archives and some commercial actors in this field. The average number of people working with web archiving issues is really low. By the half of the institutions the respondents belonging to less than one full-time professional person is focusing on this issue. About a quarter of the respondents determined the number of people between 1 and 3. Nine percent of the institutions, organisations are working with at least 10 people on web archiving activities. The other institutions and organisations employ 3-5 people for these tasks. The third question focused on the type of activities related to web archiving. Most of the related people are curating content and setting up regulations, standards. Other main tasks are (by relatively the same weight): making metadata; quality assurance, communication tasks, harvest management. The least number of people in web-archiving field are the software developers. Most of the respondents have public collection background and only a small portion of them have relevant IT experience. Most people started to work with web-archiving tasks very recently. To put these tasks to the general service portfolio of a public collection appears to be a big challenge to them. Many of the respondents referred that they are planning to work with web-archiving in the future but they do not have any practical experience recently.

The next couple of question focused on the education and training aims in web archiving field. By the answers it appears that we are still at the very beginning of our professional way. Most of the responding people recently rely on online resources in order to develop their professional competences. The number of any kind of organised training activities is marginal. A relatively large number of respondents are currently without any kind of trainings. The least number of people are attending in courses by accredited curricula. Where any kind of training option is available it mainly focusing on workshops, formulated by informal frameworks or organized by some kind of professional organisation. Most of the respondents want to develop their web-archiving related competences in IT-field, by focusing on digital preservation standards, technologies and the education of use of relevant software

[6] IIPC portal address, retrieved 12.06.2018., http://www.netpreserve.org

[7] A description of Training Working Group is available on the following link, retrieved 12.06.2018 http://netpreserve.org/about-us/working-groups/training-working-group/

[8] Survey link, retrieved 12.06.2018., https://www.surveymonkey.com/r/V7MVXXW

[9] The summary based on non-public IIPC working materials. Public references are not available.

tools. The most popular learning forms are webinars and some courses based on personal attendance.

The IIPC TWG has started to plan various training activities based on the survey experiences. The major aim is to effectively support the web archiving institutions and broad target groups of web-archiving related professionals (figure 3). The second step was the collection of a list of trainings and courses by all of their major features that are already available in various countries, and learn form the best practices. The third step has been taken recently by starting to plan an online education environment that can be used by the IIPC members in general and can be adapted to each member's needs.

Figure 3. Collaborative work in an online meeting of IIPC Training Working Group

3. Training plans in Hungary

The National Széchényi Library has started a comprehensive project in order to establish a new national library system (OKR-project). As a segment of this large project, the web archiving pilot project has started at the beginning of 2017. By consulting public collection professionals, a definite aim has appeared to establish a 30 hour long special training to this target group. The key of success of a national web archiving model is an active collaboration within the public collection sector. The main goal of the special training therefore to introduce public collection professionals to the major technical background of preserving online content, offer an overview about international projects and present our own activities in this field. By completing this course, people should be able to create web archive collections in their workplace or for private purposes. They also have to be able successfully participating on the building process of a Hungarian Internet Archive by their own competences. Main target groups are librarians, archivists, museum professionals. The course will be offered by the web archiving team of the Electronic Library Department of the National Széchényi Library and by IT professionals from the Department of IT services. *(Figure 4)*

Course curriculum consists of the following main modules:

- Getting to know internet preservation terminology, definitions and models in a basic theoretic framework.

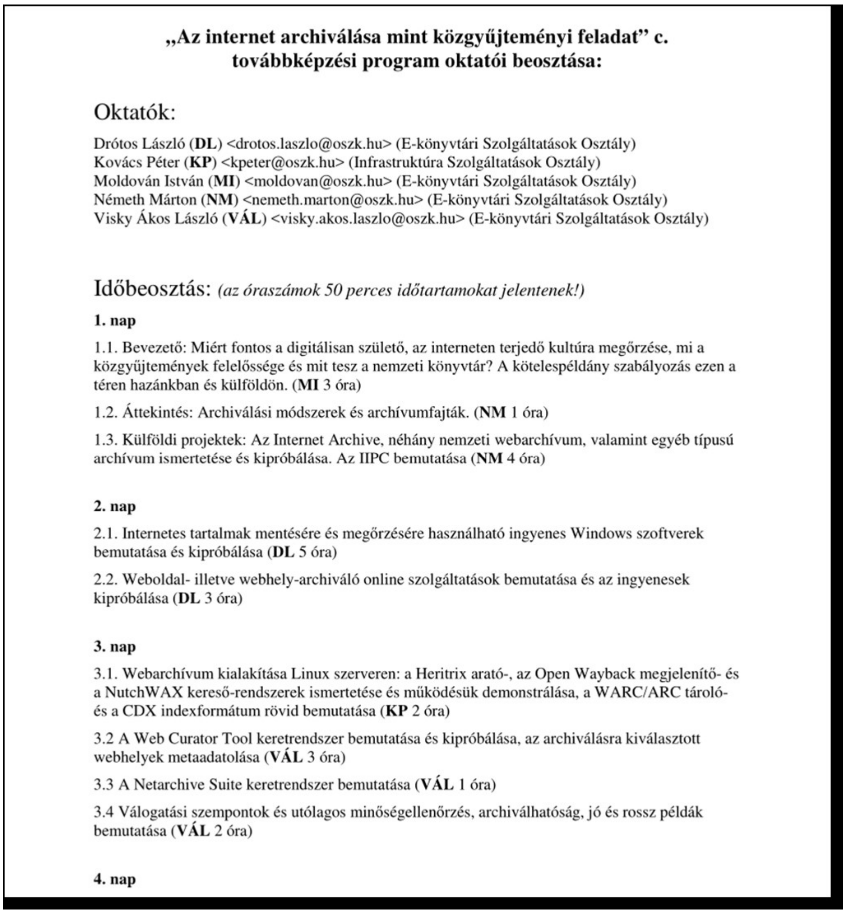
- Get competences in a basic level of using some Windows-based archiving software, online services, and other useful software tools to build-up and support an archiving workflow.

- Get basic competences in a user level to the workflow and major components of a Linux-based web archive.

- Basic competences on the curation of web materials and major tasks for metadata enrichment of the archived material.

- Introducing web archives as a research subject. A basic overview about using web archives for research purposes. Foundations of planning and managing user-centred (mainly scientific) services based on web archived materials. Major competences related to create and maintain appropriate conditions of long-term sustainability of web-archives..

Followed by the accreditation period and granting appropriate funds the course can be started at the late autumn of 2018 in our hope. Besides this course we have started to plan an online course based on a blended learning-based curriculum. Course would be completed partly online and partly by personal attendance. We hope that it also will be available at the end of 2018 by the latest for all people that are interested in long-term preservation of internet content.



„Az internet archiválása mint közgyűjteményi feladat” c.
továbbképzési program oktatói beosztása:

Oktatók:

Drótos László (**DL**) <drotos.laszlo@oszk.hu> (E-könyvtári Szolgáltatások Osztály)
Kovács Péter (**KP**) <kpeter@oszk.hu> (Infrastruktúra Szolgáltatások Osztály)
Moldován István (**MI**) <moldovan@oszk.hu> (E-könyvtári Szolgáltatások Osztály)
Németh Márton (**NM**) <nemeth.marton@oszk.hu> (E-könyvtári Szolgáltatások Osztály)
Visky Ákos László (**VÁL**) <visky.akos.laszlo@oszk.hu> (E-könyvtári Szolgáltatások Osztály)

Időbeosztás: *(az óraszámok 50 perces időtartamokat jelentenek!)*

1. nap

1.1. Bevezető: Miért fontos a digitálisan született, az interneten terjedő kultúra megőrzése, mi a közgyűjtemények felelőssége és mit tesz a nemzeti könyvtár? A kötelespéldány szabályozás ezen a téren hazánkban és külföldön. (**MI** 3 óra)

1.2. Áttekintés: Archiválási módszerek és archívumfajták. (**NM** 1 óra)

1.3. Külföldi projektek: Az Internet Archive, néhány nemzeti webarchívum, valamint egyéb típusú archívum ismertetése és kipróbálása. Az IIPC bemutatása (**NM** 4 óra)

2. nap

2.1. Internetes tartalmak mentésére és megőrzésére használható ingyenes Windows szoftverek bemutatása és kipróbálása (**DL** 5 óra)

2.2. Weboldal- illetve webhely-archiváló online szolgáltatások bemutatása és az ingyenesek kipróbálása (**DL** 3 óra)

3. nap

3.1. Webarchívum kialakítása Linux szerveren: a Heritrix arató-, az Open Wayback megjelenítő- és a NutchWAX kereső-rendszerek ismertetése és működésük demonstrálása, a WARC/ARC tároló- és a CDX indexformátum rövid bemutatása (**KP** 2 óra)

3.2 A Web Curator Tool keretrendszer bemutatása és kipróbálása, az archiválásra kiválasztott webhelyek metaadatolása (**VÁL** 3 óra)

3.3 A Netarchive Suite keretrendszer bemutatása (**VÁL** 1 óra)

3.4 Válogatási szempontok és utólagos minőségellenőrzés, archiválhatóság, jó és rossz példák bemutatása (**VÁL** 2 óra)

4. nap

Figure 4. An excerpt from the preliminary course-plan (in Hungarian)

Epilogue

In our paper an overview has offered about the structure and outcomes of an online professional course about web archiving that has managed from Denmark. We also offered a summary about the preliminary plans and basic activities of the IIPC Training and Working Group that offered us a major overview about the current framework, background and status of web archiving activities throughout the world. Last but not least we elaborated our training plans in Hungary. It is vital to train people with certain competences in order to build-up a national web-archive network. Based on this collaborative framework archiving activities can be done ordinarily and efficiently. A major pre-condition of the establishment of a well-functioning national network is to guarantee permanent professional development (both individually and on institutional level). Accredited trainings must be offered for web-archiving professionals in a permanent way to constantly keep their knowledge on a required level.

Bibliography

IIPC Training Survey Call, retrieved: 12.06.2018
https://netpreserveblog.wordpress.com/2017/12/14/iipc-training-survey/

IIPC Training Survey, retrieved: 12.06.2018
https://www.surveymonkey.com/r/V7MVXXW

IIPC Training Working Group portal, retrieved: 12.06.2018
http://netpreserve.org/about-us/working-groups/training-working-group/

NIELSEN Janne, *Using the Web archives in Research (Theoretical course book of NetLab web archiving course),* retrieved: 12.06.2018
http://netlab.dk/wp-content/uploads/2016/10/Nielsen_Using_Web_Archives_in_Research.pdf

Website of the Danish Web Archive, retrieved: 12.06.2018
http://www.netarkivet.dk

A brochure of NetLab web archiving course retrieved: 12.06.2018
http://netlab.dk/wp-content/uploads/2017/04/NetLab-Web-Archiving-Course-Brochure.pdf

Website of the NETLAB web archiving course, retrieved: 12.06.2018
http://netlab.dk/services/courses/