Преглед НЦД 16 (2010), 18–24

Bogdan Trifunović
Public Library "Vladislav Petković Dis"
Čačak, Serbia

WEB ARCHIVING PROJECTS END-USER PERSPECTIVE

Abstract. This paper examines usability and accessibility of the publicly opened Web archiving projects. It aims to identify user-friendly features associated with the web sites of several web archiving projects, but also the creation of basic structure and framework for the comparative analysis in the process of evaluating them. For the purpose of this research, five Web projects have been selected: PANDORA (National Library of Australia), EUROPEAN ARCHIVE, MINERVA (Library of Congress), UK WAC and WEBARCHIV (National Library of the Czech Republic).

Projects were "inspected"[1] from the perspective of the regular user, concentrated on ease of navigating the web sites, accessibility (e. g. position of search filed, search options, browsable categories, organization of the content, help section), documentation and general layout (analysis of implemented interface design and it's looks, where following modern trends in web design is not regarded as crucial advantage over old school in design and layout). The stress is on user-oriented experience of using these web sites and their background architecture. The ultimate goal is to raise awareness in Serbia and the Balkans about Web Archiving and digital preservation of the Internet resources and online cultural heritage.

Key words: web archiving, Internet, digital preservation, review, users.

1. Web Archiving Overview

The World Wide Web, or just the Web, is generally speaking the "Internet" how we are used to call it. The immense number of documents types is what the Web consists of. The Web, at the same time, is content and medium, so it must be observed in two dimensions.[2] Web archiving stands for the aim to capture (harvest), store (archive) and access (open) online documents in the offline world. Web archiving projects major challenge ariuse from dynamic and ephemeral nature of the Internet.[3] Therefor, web archiving is a part of more complex and thorough process of digital preservation, as a "range of activities required to ensure that digital objects remain accessible for as long

---

[1] The review process of the Web archiving projects was conducted in November 2008, in the National Library of the Czech Republic. This research was made possible by DigitalPreservationEurope Exchange Programme (DPEX) http://www.digitalpreservationeurope.eu/exchange/. The author expresses his sincere gratitude to all colleagues at DPEX and National Library of the Czech Republic in Prague (NKP, http://www.nkp.cz/), where an exchange visit was accomplished in November 2008. The first review results were presented at the Eighth National Conference *New Technologies and Standards: Digitization of National Heritage 2009*, organized by the National Center for Digitization in Belgrade, April 15-16, 2009. The information in this article represents the views of the author alone and does not necessarily reflect the views of the Public Library Čačak, National Center for Digitization, DPE or the National Library of the Czech Republic.

[2] Julien Masanes (ed.), *Web Archiving*, Springer, 2006, 55.

[3] "Web archiving initiatives exist to collect ephemeral Web content for use by current and future generations of users." (Masanes, 177).

as they are needed".[4] It is also the only way of keeping the past and present cultural heritage on the Internet for the future, because the Web content is growing every day and more content is only represented in a digital form. Some strategy for long-term archiving method of that material must be utilized in an effort to overcome the possibility of loosing our heritage, because society didn't realize its historical value on time.[5] Projects aiming to "capture, store and make it [Web domain] accessible for the long term" are confronted usually with the non-existence of the proper legislation, which would treat digital publications as printed and therefore extend legal deposit to digital publications. In some countries, like Canada, Denmark, New Zealand, Norway, the UK or the Czech Republic this process is already enacted, but many more still have to wait for necessary changes in the law.[6]

National libraries in Austria, the Czech Republic, Denmark, Finland, France, the Netherlands, Norway, Sweden and the UK, have started to build national web archives (repositories) using a variety of approaches: 1. selective archiving of static and/or dynamic web resources; 2. whole of domain harvesting (automatic capture of the whole national domain); 3. a combination of the selective and whole of domain harvesting (e. g. a resources of high research value).[7] For instance, the National Library of the Czech Republic in Prague since 2000 preserves "electronic online publications", as a part of the WebArchiv[8] project, which is a digital archive of Czech web. The sheer size of the Web and the enormous number of online documents, but also the quality of the Internet publications, impose on the authors of WebArchiv selection criteria for including online resources in the digital archive. The aim is that documents of a current and future value which constitute the Czech national cultural heritage on the Internet are preserved.[9] The National Library of the Czech Republic accepted twofold criteria for web archiving, based on previous experience: comprehensive archiving of web resources and selective approach (harvesting of resources for which the written permission-contract for online access to the archived copies was obtained from the publishers of the resources). This approach to web archiving divides the criteria into two categories depending on the method of the acquisition or rather the legal conditions of providing access to the archived data.[10]

The oldest web archiving project is Internet Archive (IA)[11], established in 1996 as a non-profit organization. IA uses Alexa[12] robot (crawler) to archive web sites, but also Heritrix[13] crawler (built in 2003) to create the snapshots of the entire WWW (until now more than 150 billion web pages). Although IA claims that it is capturing the whole Web, the size and scope of the Internet is making that task almost impossible. For instance, the Public Library in Čačak web site was captured 145 times from January 28, 2001 to January 15,

---

[4] Masanes, 178. More on digital preservation in: Henry M. Gladney, *Preserving Digital Information*, Springer 2007.
[5] Peter Lyman, *Archiving the World Wide Web*, in: Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving, CLIR-Library of Congress, April 2002, 38-39, acessed online at http://www.clir.org/PUBS/reports/pub106/pub106.pdf#page=42 on August 15, 2009.
[6] Digital Preservation. *Calimera Guidelines*, accessed online at http://www.calimera.org/Lists/Guidelines/Digital_preservation.htm on 31 August 2009.
[7] Ibid.
[8] WebArchiv, http://www.webarchiv.cz/, accessed online on 15 May 2009.
[9] WebArchiv – Archive of the Czech Web: Selection criteria for web resources, accessed online at http://en.webarchiv.cz/criteria/ on 31 August 2009.
[10] Ibid.
[11] Internet Archive, http://www.archive.org/index.php, accessed online on 2 September 2009.
[12] Alexa, http://www.alexa.com/, accessed online on 2 September 2009.
[13] Heritrix, http://crawler.archive.org/, accessed online on 2 September 2009.

2008[14], but that was also the last capture for the past 18 months. In the meantime the national domain of Serbia was changed from .yu to .rs, which affects our web site address, but IA has not made harvest of it yet. In the more recent time the International Internet Preservation Consortium (IIPC) was formed, as international consortia of institutions world wide for the collaborative development of a set of tools for web archiving and accessing collections.[15]

Mentioned obstacles in the process of web archiving the whole Web created newer approaches, mostly dealing with the "national" part of the Internet (e.g. capturing and archiving national domain), where archiving of the online documents could be seen as a digital preservation of "the Web heritage". These projects are usually run by major national institutions (libraries or library consortia) and mostly concentrate on selective approach of identifying quality content on the Internet for long-term archiving.[16]

2. Web Archiving Projects Review

Projects included in the review:

- PANDORA http://pandora.nla.gov.au (National Library of Australia), harvesting and archiving Australian web domain
- EUROPEAN ARCHIVE http://www.europarchive.org (non-profit foundation), digital library of cultural artifacts in digital form
- MINERVA http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html (Library of Congress, Washington), web archived thematic collections regarding the USA
- UK WEB ARCHIVING CONSORTIUM http://www.webarchive.org.uk (a consortium of six UK institutions), harvesting and archiving UK domain
- WEBARCHIV http://www.webarchiv.cz (National Library of the Czech Republic), digital archive of Czech web resources

The scope of the review: analyzing usability and accessibility of the publicly opened web archiving projects.

The aim of the review: identifying user-friendly features associated with the above mentioned projects web sites, but also the creation of basic structure and framework for comparative analyses in the process of evaluating them.

2.1. Criteria and focus. For the purpose of this review we selected five web archiving projects accessible online for the users: PANDORA, EUROPEAN ARCHIVE, MINERVA, UK WEB ARCHIVING CONSORTIUM and WEB ARCHIV. The criteria were that a project (or some part of it) should be open to general public over the Internet and that it has to show some archived materials. Thus, national projects like in Sweden or Denmark were not considered for this research, because they are not open for the Internet users. Every project web site was "inspected" from the perspective of the regular user, concentrated on the ease of navigating the web site, accessibility (e. g. position of search filed, search options, browsable categories, organization of content, help section), documentation and general layout (analyzes of implemented interface design, it looks and "modernity", where following the modern trends in web design is not regarded as crucial advantage over old school in design and layout). The stress was on the user-oriented experience of using these web sites and their background architecture. The technical aspect was left aside in this case.

[14] http://web.archive.org/web/*/http://www.cacak-dis.org.yu

[15] The IIPC's members are more than 30 national and regional institutions world wide (libraries and archives). More on IIPC on their web site http://www.netpreserve.org/about/index.php.

[16] "The selection phase is a key phase in Web archiving." (Masanes, 71).

2.2. Interface design. General looks and feels of reviewed projects vary a lot, which was expected. That's mostly the result of the organization of the web sites' layout. Most web sites look fashioned in the old way and manner, with only one implementing some new design trends and Web 2.0 elements (tag cloud, My Desktop) – European Archive. It must be said that European Archive web site is the newest, so it could be expected that eventual redesign of others would bring similar change. WebArchiv web site is somewhere in the middle between old web design and modern one. UK WAC web site is currently in reconstruction (which left impact in its functionality too). UK WAC system is a clone of PANDORA project, using PANDORA Digital Archiving System (PANDAS), but the creators of that web site also copied the layout of Australian project.

If we left aside impression on design and look, which is quite disputable argument in evaluating the Internet projects (but more and more pressuring in the course of time), the significant values would be usability and accessibility. We should also discuss the quality of help sections for users, project documentation and search options.

2.3. Usability. Most of projects proved to be quite manageable in the sense of navigation through web site pages, browsing the content or collections. It depends, of course, on the amount of content, where navigating and browsing the content of European Archive web site was very easy, counting the fact that there isn't a lot of web content anyway. PANDORA and MINERVA projects also implemented understandable and usable layout, without sufficient elements to distract users from the main content. WebArchiv project's web site could distract users with slightly confusing navigation and links for browsing collection on the bottom of the navigation tree. As only non-English project considered here, with the main interface in languages other than English, it must be taken into account that the Czech interface (the main one) is slightly richer in content and options than the other one, but the remarks on navigation and browsing stay. But, on the other side, WebArchiv provides information on the web site home page of the last modification and up-to-date data stored in the system. Only MINERVA provides a date of last modification on its home page.

2.4. Accessibility. Accessibility plays maybe the most important part in the final decision on usability of reviewed web sites. For the purposes of this paper we concentrated on the access to content and the ways users could browse content (collections). Unlimited access to the all provided content online is in PANDORA, European Archive, UK WAC and WebArchiv projects. Users are able to see all of represented collections and access archived web sites. PANDORA web site provides browsing of the content by subject and by title, which is similarly implemented in UK WAC web site, with addition of browsing thematic collections in UK WAC. European Archive provides multilingual interface (the only one, WebArchiv is bilingual) and it is the only digital preservation web project in full sense among these five, where web archiving is part of the project. That is why European Archive has its content divided by the type of resource, but all provided materials are fully accessible and browsable. WebArchiv also implements unlimited access to "contracted" web sites, which is the database of web sites with written contracts from the publishers of the web resources, where they agreed with public access to their archived resources. At the time of this review (end of November 2008) there were 737 fully accessible web sites. On the Czech version of the WebArchiv (the main interface) users are able to browse collections by subject and title.

On the other hand, MINERVA has restrictions on using and accessing content of the some of its thematic collections, which could only be access from the Library of Congress. The content of accessible web sites can be browse by subject, title and (personal) name.

2.5. Search options. In the background of all analyzed projects lies a database. As every database, its full usefulness could be reached only through records search option. Browsing 737 web sites inside WebArchiv project could be labor intensive and time consumptive task, but searching all those records in some way will produce results much quicker. As for WebArchiv, its database is searchable only by URL address of "contracted" web sites, which put users in an awkward situation, that they should browse collections first and collect data, to be able to use the search option. That suggests that content of archived web sites inside WebArchiv isn't indexed. The search by URL is usable only if user knows in advance the exact URL address he or she is looking for.

Probably the best search option is provided in PANDORA project. Users are able to conduct basic (term/s) search, but also advanced search of content and URL, with choosing the number of search results to be displayed, limiting search query by the subject categories or by the date of archiving. Boolean operators such as AND, "+", OR, NOT and "-" are supported, wild card searches with "*" may be used, as well as search by more complex phrases. There is also well documented search help, which helps users with the search options in PANDORA, providing them all necessary information needed for the basic and advanced search.

UK WAC project web site has only basic search function of the content, but at the time of this review it didn't work. Similarly, European Archive also uses only basic search for its collections, but that functionality is not implemented for the web content, so users could only browse the collections.

MINERVA project implements basic search of its collections, which could be limited on particular metadata (name, title, subject, language, etc.) or collection. It is possible to refine search with Boolean operators. The search help section is also presented and it provides users with all info they needed.

2.6. Project documentation and help sections. The project documentation proves to be challenging task for the project management, especially if the creation of documentation didn't follow the workflow and the project's phases. Considering these five projects of web archiving, three of them (60 percent) come with poor or inadequate documentation about the projects, their aim, scope, activities, used technologies, user-oriented information etc. MINERVA project leads in its scanty details about the technical aspects of the project and nothing more. UK WAC provides more info in the section about the project, but still it is inadequate considering the proposed aim of this project and its broadness. European Archive stands the best among the three, with fairly good documentation on web archiving, digitization, used infrastructure for the project, supporting institutions, and detailed Terms, Privacy & Copyright page.

PANDORA and WebArchiv projects use a different approach, documenting much more and providing interested parties (general users, scholars, colleagues, institutions, government officials) all the information they could use. PANDORA has excellent and very detailed documentation, which sometimes goes much broader in topic and regards issues dealing with digital preservation in general. Just section "About Pandora" could be the foundation of a new website, with overview of the project, its history, policy and practices, selection criteria, manuals, software platform, staff papers, legal deposits etc. Besides that, there is also data about project statistics, information on services, disclaimer etc.

Another good example on documenting current work is WebArchiv, with detailed overview of the project, aim, access, standards, included staff papers, presentation and other articles about WebArchiv, links to relevant resources and projects, etc.

Regarding previously said about PANDORA and WebArchiv projects, it is clear that they are friendlier web environments than the other three, where users may collect all necessary information at one place. For instance, PANDORA also provides FAQ section, with more than 25 questions and answers. All web sites have contact information, varying from pages with contact forms in MINERVA and UK WAC projects, or only one email address in European Archive (it is actually a link to email address, which starts users email client on the computer after clicking on Contact, an outdated approach avoided by most users today).

2.7. Conclusion. Web archiving projects are common feature today, but significant number of them is closed for the general public (mostly because of legal obstacles). Analyzed web sites in this review are selection of well known and established projects, open to the Internet users worldwide. Some conclusions on their usability and accessibility derive from analyzed elements. PANDORA project stands near the top among the five, simply beating others in some key elements, as search options, help section, documentation or usability. But we must not forget that some of the projects exist longer than the others, or that some of them are in the process of reconstruction, so that could be also taken into account. This review is not trying to give decisive conclusions, or to represent highest authority for the topic, but it is one way of considering positive and negative aspects of web archiving projects. In that sense, PANDORA project is really paying attention on usability and user-friendly features, while others are good in some aspects but not in all of them. WebArchiv proved to be highly potential project, while European Archive needs more content for the proper evaluation. MINERVA and UK WAC are slightly outdated in their general appearance, but also in their usability and final usefulness, which should be changed with suggested reconstruction[17] of UK WAC project and future work on web archiving at the Library of Congress and other institutions.

References

1. Julien Masanes (ed.), *Web Archiving*, Springer, 2006.
2. Henry M. Gladney, *Preserving Digital Information*, Springer, 2007.
3. Peter Lyman, *Archiving the World Wide Web*, in: *Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving*, CLIR-Library of Congress, April 2002, http://www.clir.org/PUBS/reports/pub106/pub106.pdf#page=42
4. Digital Preservation. *Calimera Guidelines*, http://www.calimera.org/Lists/Guidelines/Digital_preservation.htm
5. WebArchiv, http://www.webarchiv.cz/
6. WebArchiv – Archive of the Czech Web: Selection criteria for web resources, http://en.webarchiv.cz/criteria/
7. Internet Archive, http://www.archive.org/index.php
8. Alexa, http://www.alexa.com/
9. Heritrix, http://crawler.archive.org/
10. International Internet Preservation Consortium, http://www.netpreserve.org/about/index.php

[17] UK WAC web site has been reconstructed in the meantime.

Bogdan Trifunovic

WEB ARCHIVING PROJECTS USER-ORIENTED COMPARATIVE REVIEW - November 2008

Web Archiving project	Project documentation	Interface design	Usability	Accessibility	Search options	Web Crawler
PANDORA pandora.nla.gov.au	Excellent, very detailed and broad	Simple and descriptive, old web design	Very easy to navigate and browse the content	Unlimited access, web content browsable by subjects and by title	Basic and advanced search by content and URL, Boolean op, search help, limiting search results	HTTrack
EUROPEANARCHIVE * www.europarchive.org	Poor documentation	Modern and descriptive, some Web 2.0 elements (tag cloud)	Quite easy to manage through collections (though there is no lot of content)	Unlimited access, multilingual interface, content divided by movies, recordings and web	Basic search option, but not implemented for web content	Heritrix
MINERVA http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html	Poor documentation	Descrptive, old web design	Easy to navigate and browse the content	Restrictions on some collections, content browsable by subject, name and title	Basic search option, could be limited on particular metadata or collection, Boolean, search help	Crawl by Internet Archive (Heritrix)
UK WAC ** www.webarchive.org.uk	Poor documentation	Simplified PANDORA clone, old web design	Easy to navigate and browse the content	Unlimited access, content browsable by subject, thematic collections and title	Only basic search option (not working at the time)	HTTrack
WEBARCHIV www.webarchiv.cz	Good documentation	Fairly modern and descriptive design	Little bit confusing navigation, browsing collections should be more emphasized	Unlimited access to, content browsable by collections and title (Czech version)	Basic search only by URL address of "contracted" web sites	Heritrix

* Web site claims that project is still in development

** PANDORA technology implementation, web site currently in reconstruction

bogdan@cacak-dis.rs
www.cacak-dis.rs