

Bogdan Trifunovic, M. A.

Information Manager

Public Library Cacak

Serbia

www.cacak-dis.rs

WEB ARCHIVING PROJECTS USER-ORIENTED 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

Date of review: November, 2008

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

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

* See page 6 for disclaimer and acknowledgment.

CRITERIA AND FOCUS

For the purpose of this review we selected five web archiving projects accessible via internet for the users: PANDORA, EUROPEAN ARCHIVE, MINERVA, UK WEB ARCHIVING CONSORTIUM and WEB ARCHIV. The criteria was that a project (or some part of it) should be open to general public over 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 internet users.

Every project web site was “inspected” from the perspective of the regular user, concentrated on ease of navigating the web site, accessibility (e. g. position of search field, search options, browsable categories, organization of content, help section), documentation and general layout (analyzes of implemented interface design, it looks and “modernity”, 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 technical aspect was left aside in this case.

INTERFACE DESIGN

As expected, general looks and feels vary a lot. That’s mostly the result of organization of the layout of the web sites. Most web sites look fashioned in 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 valuing 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, search options, etc.

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 implement 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 navigation tree. As only non-English project considered here, with a main interface in other language than English, it must be taken into account that 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 last modification and up-to-date data stored in the system. Only MINERVA provides a date of last modification on its home page.

ACCESSIBILITY

Accessibility plays maybe the most important part in final decision on usability of reviewed web sites. For the purposes of this paper we'll concentrate on access to the content and the ways users could browse content (collections). Unlimited access to all provided content online is provided 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 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 (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 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.

SEARCH OPTIONS

In the background of all analyzed projects lays a database. As every database, its full usefulness could be reached only through records search options. Browsing 737 web sites inside WebArchiv 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 awkward situation, that they should browse first collections and collect data, to be able to use search option. That suggests that content of archived web sites inside WebArchiv isn’t indexed. URL search is usable only if user knows in advance the exact URL address he or she is looking for. Probably best search option is provide 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 subject categories or by 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 search options in PANDORA, providing them all necessary information needed for 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 web content, so users could only browse 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. Search help section is also presented and it provides users with all info needed.

PROJECT DOCUMENTATION AND HELP SECTIONS

Project documentation proves to be a challenging task for project management, especially if the creation of documentation didn't follow the workflow and 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 technical aspects of the project and nothing more. UK WAC provides more info in 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 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 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 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, which is outdated approach most users avoiding today).

CONCLUSION

Web archiving projects are common feature today, but significant number of them is closed for general public (mostly because of legal obstacles). Analyzed web sites in this review are selection of well known and established projects, open to 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 others, or that some of them are in the process of reconstruction, so it 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 features and content for 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 of UK WAC project and future work on web archiving at Library of Congress and other institutions.

*Disclaimer: The information in this article represents the views of the author alone and does not necessarily reflect the views of DPE or the National Library of the Czech Republic.

*Acknowledgment

This research was made possible by DigitalPreservationEurope Exchange Programme (DPEX) <http://www.digitalpreservationeurope.eu/exchange/>. The author express sincere gratitude to all colleagues at DPEX and National Library of the Czech Republic in Prague (NKP, <http://www.nkp.cz/>), where exchange visit was accomplished in November 2008.

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	Descriptive, 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