Web Archiving as a Service for the Sciences

Anna Kugler
Munich Digitization Center/Digital Library
Bavarian State Library
anna.kugler@bsb-muenchen.de

Tobias Beinert
Munich Digitization Center/Digital Library
Bavarian State Library
tobias.beinert@bsb-muenchen.de

Astrid Schoger
Munich Digitization Center/Digital Library
Bavarian State Library
astrid.schoger@bsb-muenchen.de

ABSTRACT

The collection and archiving of scientifically relevant websites is so far a vastly neglected sphere of activity in German libraries. To counteract this looming loss and providing researchers with permanent access to websites, the Bavarian State Library (BSB) has built up a web archiving workflow already more than two years ago. The main goal of a project newly approved by the German Research Foundation (DFG) is the development and implementation of a cooperative service model. This service will support other cultural heritage institutions in their web archiving activities and facilitate the build up of a distributed German scientific web archive. With this project the Bavarian State Library wants to improve both quantity and quality of scientific web archives and promote their use in the scholarly context.

Keywords

Digital Preservation, Web Archiving, Web Harvesting, Bavarian State Library

1. INTRODUCTION

Digital communication will co-determine the future of the humanities. This statement was beyond all questions at the conference 'Reviewing – Commenting – Blogging: How will Humanities Scholars Communicate in the Digital Future?', which was organized by the Bavarian State Library (http://www.bsb-muenchen.de) at the beginning of this year. The conference was held on the occasion of the 2nd anniversary of the review platform for European History named recensio.net where the participants discussed how far the web as a publication platform can meet the qualitative requirements of scholarship [1]. The presentations and round table discussions clearly showed that new instruments, services and infrastructures need to be developed to publish, evaluate and finally preserve these new types of scholarly resources. One of the new digital resources which have often been "almost completely overlooked", as Nils Brügger put it, "even if they may be considered one of the most significant contemporary contributions to the cultural heritage of mankind" [2] are websites.

Web archives which preserve captures of websites and make them permanently available are still an unknown or unusual type of research instrument for many researchers. Compared to the live web a few distinctive features of web archives however exist, which constitute their necessity:

1. Web archives include content which has already disappeared from the live web. The estimates about the average lifespan of websites differ a lot, but what they show nevertheless is that "although the web can be considered a storage medium of our civilisation, it does not preserve itself for the future – the old web cannot always be found on the web." [3]

2. Already today one concrete use is the possibility to cite websites. Scientists more often refer to or cite online resources, but disappearing content or changing URLs often make consistent access to the cited sources quite difficult or even impossible.

3. Moreover the history of the web illustrates an important part of our culture. Periodic captures of websites not only show the evolution and changing of web technology and web design but also the changing of political and scholarly discourses.

4. In a more technical context it could be possible in the future that certain documents cannot be separated from the tools or platforms which produced them. Based on this fact archiving of websites has a totally different use than archiving of printed books and at the same time makes it much more challenging [4].

5. Last but not least web archives offer a subject-oriented data collection which can be analysed by new types of data mining methods. In the context of the emerging e-humanities scientists can be offered advanced access possibilities.

Nevertheless experience reports of web archives already operating for a long time, such as e.g. the UK Web Archive show that there is still "little evidence of scholarly use" [5]. Next to the fact that many scholars don't know about the existence of web archives, already active users would appreciate an increase of scientific content and see a need for improved data mining tools [6]. Thus the aim of the whole web archiving community, which mainly consists of national libraries and larger regional libraries, has to be to improve quantity and quality of scientific web archives and to promote their use in the scholarly context [7].

2. A DISTRIBUTED GERMAN RESEARCH LIBRARY – RESULTING IN A DISTRIBUTED GERMAN RESEARCH WEB ARCHIVE?

In Germany the state of web archiving activities looks a little bit like a rag rug of small initiatives collecting websites. In some cases the collection strategy derives from a regional legal deposit obligation (e.g. BOA, Baden-Württembergisches Online-Archiv, http://www.boa-bw.de/ or edoweb, digital archive for Rheinland-Pfalz, http://www.lbz-rlp.de/cms/landeskunde/edoweb/), while other regional libraries do not archive websites at all. The German National Library, which is the legal deposit library for Germany, started harvesting the most important websites significant for German society, history, politics, economics and culture last year and intends to make them accessible in their reading rooms [8]. In summary web archiving in German libraries does not fulfil the needs of the scientific community so far. This is partially owed to the federally structured German library landscape.

In Germany due to the political and historical situation one central research library has never been established. In 1949 a special cooperation system was launched which is internationally unique and takes responsibility for the supra-regional literature supply of scholarship and research. It is organised by the German Research Foundation. 23 state and university libraries and some specialised libraries participate in this cooperation system, each of which is responsible for one or more scientific subjects (Sondersammelgebiete, SSGs). All libraries participating in this system pledge themselves to make their collections available nationwide. The access point to the collections is offered by so-called virtual subject libraries which bundle the diverse search possibilities under one user interface („one-stop-shop“) [9].

The collections of these subject libraries include freely accessible internet resources with scientific relevance which are to a great extent websites. Each website is intellectually selected by experts of the respective specialist department and a lot of effort is put into the cataloguing process. Several libraries decided to cooperate concerning the cataloguing system and they built up Academic Link Share (ALS) (http://www.academic-linkshare.de), a database system already containing more than 100,000 entries. Not only the content itself but also this time-consuming work is lost as soon as the respective website disappears, as until now archiving is not mandatory for these subject libraries.

A thorough evaluation process of the system of special subject areas commissioned by the German Research Foundation in 2010 with the aim to find out how the needs of the sciences could be best fulfilled in the future resulted in several recommendations: a future acquisition focus on electronic resources, a stronger focus on customisation to the scientific needs concerning collection building, and in general an improved service-orientation for scholarship. In order to guarantee sustainability, long-term preservation and permanent access to digital data new service models should be developed [10].

In this context BSB applied for a project at the DFG which focuses on developing a cooperative service model for web archiving. It will be built on the infrastructure and experience BSB has already gained in creating a scientific web archive over the last two years and thus support other cultural heritage institutions in their web archiving activities. A closer cooperation with scientific communities is an important aim. The outcome of the project could become the nucleus for a distributed German Research Web Archive. The project started at the beginning of 2013.

3. WEB ARCHIVING AT THE BAVARIAN STATE LIBRARY

In 2010 BSB's Munich Digitization Center/Digital Library (MDZ) (http://www.digitale-sammlungen.de) began to collect and archive websites in a pilot phase, since the beginning of 2012 its web archive increases productively (http://www.babs-muenchen.de). The focus is on websites already selected and catalogued by the virtual subject libraries (Virtuelle Fachbibliotheken, ViFas). BSB e.g. is responsible for five virtual subject libraries dealing with the fields of:

- History (www.propylaeum.de, www.historicum.net)
- Musicology (www.vifamusik.de)
- Eastern Europe (www.vifaost.de)
- Romanic culture area (www.vifarom.de)
- Library, book and information studies (www.b2i.de)

These virtual subject libraries account for about 10,000 entries of websites in the ALS database. As Bavaria's existing legal deposit regulations only allow to harvest and archive websites of Bavarian authorities but not scholarly websites in an international context an explicit permission is necessary to harvest, archive and make accessible those websites. Thus a very detailed permission request (email) is sent to each 'website owner' in which the rights to harvest and archive the website in regular sequences and to make it available on the web are requested if no rights of third parties interfere. Moreover in the permission request email it is pointed out that German copyright law applies. The positive return rate is about 20% to 30%.

For the harvesting process BSB uses the Web Curator Tool (WCT), an open source software developed jointly by the British Library and the National Library of New Zealand. It was chosen because it allows an integrated process for selective web harvesting including the administration of the permission request, harvesting with job scheduling and a partly automated quality control. The website crawler is Heritrix which has been developed by the Internet Archive. To provide access to the crawled websites a local adaptation of the Wayback Machine has been implemented. The harvested WARC (Web ARChive) files are archived in Rosetta, a commercial software for digital long-term preservation by Ex Libris. Ensuring bitstream preservation is done by the Leibniz Supercomputing Centre (LRZ), whom the BSB has been working together closely for many years now.

All harvested and archived websites are freely accessible. The harvested websites are made available firstly via the index of the virtual subject libraries and secondly via our local library catalogue. The first access point to the archived websites is based on the indexed metadata of the virtual subject libraries. For each internet resource there is a detailed ALS entry containing title, URL, responsible institution, contact, key words and even a short abstract. Next to the link to the live website there has been added a so-called „Archive-URL“, which is a stable link leading to the archiving system Rosetta, where the websites are preserved. If the user clicks on this “Archive-URL” link a local adaptation of the Wayback Machine opens and shows all the archived captures for a single website. For citation purposes persistent identifiers (Uniform Resource Names, URNs) will be assigned on website level or even on capture level. In BSB’s catalogue system each website is described by a minimal catalogue entry. About 500 archived websites (with several captures) can be found in BSB’s catalogue system so far (June 2013).

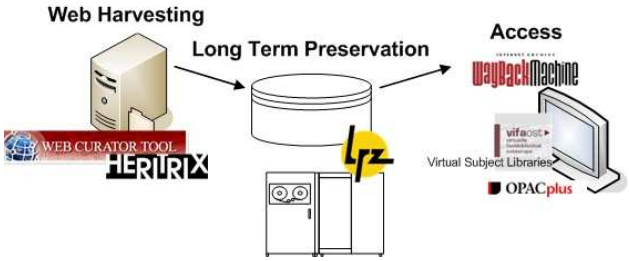

Figure 1: BSB’s Web Archiving Workflow

4. BUILDING UP A SERVICE FOR OTHER CULTURAL HERITAGE INSTITUTIONS

The newly approved DFG project focuses on creating a web archiving service for other (German) cultural heritage institutions like e.g. the Web Archiving Service (WAS) of the Californian Digital Library (http://webarchives.cdlib.org) or Archive-It from the Internet Archive (http://www.archive-it.org). Building up this service is accompanied and based on several improvements and extensions of the existing infrastructure.

The first work package of the project aims at developing a collection and archiving profile for websites which will constitute a very important guideline for the institutions using the new web archiving service. The collection profile will rely a lot on the collection criteria for scientific internet resources already defined by DFG, but additionally include criteria which derive from BSB’s specific collection policy as archiving library of Bavaria. Very challenging is the definition of an archiving profile which has to address content criteria as well as technical and temporal coherence aspects. Due to technical challenges the available harvesters often are not able to crawl an identical mirror of a website, especially dynamic functionalities like e.g. database queries that won’t work any longer in a web archive. Maybe these URLs need to be excluded from the archiving process although the content is highly valued due to the collection profile. Moreover the crawl duration and frequency is of course critical referring to completeness and temporal coherence of a harvested website [11]. Thus an archived website can always only be a “reconstruction” [2] of the live website, which needs to be reflected in the archiving profile. A detailed collection and archiving profile will help to design the defined crawl scope more exactly and to improve the quality of the web archive.

Improving the quality assurance process and defining quality criteria is another task in the new project, which still poses a challenge for many institutions in the web archiving context [12]. At BSB the quality of a new website capture is still evaluated manually by visually comparing the live website with the archived version and by analysing the log files of the crawler. But the increasing amount of website captures makes the development of more automated workflows inevitable. Thus defining quality criteria and testing new tools and methods already used by other institutions and recommended by the International Internet Preservation Consortium (IIPC, http://www.netpreserve.de) will help to develop a (partly) automated workflow. Moreover the new version 1.6 of the WCT offers a partially automated workflow for quality assurance on the basis of technically reviewable quality indicators, which will be implemented as soon as possible.

Also in terms of access new ways have to be explored, as a mere searching for URLs or titles of websites in many cases no longer fulfils researchers’ needs. [13]. Therefore the already existing data mining technologies, that might be appropriate for web archives have to be analysed and tested. A special focus is to be put on full text indexing and search, with SOLR being considered as a possible solution [14]. Another way to improve the accessibility of the content of the web archive could be to work on Memento compatibility, so that historical versions of websites from different web archives can be integrated into the live web via a simple browser extension [15].

Another work package deals with available long-term preservation measures for web archives. Different tools and methods for the use of digital long-term preservation of websites will be tested and probably implemented like e.g. JHOVE2 for the format characterisation of the files inside the WARC files. On the basis of this format characterisation several format migration tests will be pursued. Moreover deduplication tests are part of this work package in order to gain more knowledge about the possibilities of storage reduction processes with the Tivoli Storage Manager (TSM) which is used at the LRZ or already at harvesting time with the latest Heritrix version. The preservation system Rosetta offers certain risk assessment functionalities and format migration solutions which will be tested for websites during the project.

Based on these experiences and improvements the conceptual phase will start for building up a cooperative service model for web archiving for other (German) cultural heritage institutions. To achieve this BSB will work together with external partners, find out about their requirements for web archiving and design a basic service concept with different service levels. First of all the service levels cover the selection and harvesting process which could be done centralised at BSB or decentralised with a complete new software installation at the partner institution. Secondly the archiving and preservation responsibilities need to be discussed which will be most probably solved best centralised at BSB. The third service level deals with access, which depends very much on the requirements of the partner institution. The most crucial point

is whether free access is possible or not and which search possibilities are favoured.

After the conceptual work is done, a technical implementation of the service model has to be realized. That includes most probably further installations of the WCT, setting up the required interfaces and possibly own technical extensions. In a final step after an intensive testing phase the overall costs for a cooperative web archiving routine have to be precisely calculated and put into a business model.

5. CONCLUSION

The ambitious work programme described above aims at improving the state of web archiving in Germany. With an extended and cooperative infrastructure the already existing selection and cataloguing capacities can be integrated in a much more sustainable process that includes an archival component. Thus enduring access to scientifically relevant websites for researchers can hopefully become the norm rather than the exception.

6. REFERENCES

[1] http://rkb.hypotheses.org/410and http://www.ahf-muenchen.de/Tagungsberichte/Berichte/pdf/2013/038-13.pdf

[2] Brügger, N. & Finnemann, N. O. 2013. The Web and Digital Humanities: Theoretical and Methodological Concerns. In: Journal of Broadcasting & Electronic Media 57,1 (2013), p. 66-80, here p. 79.

[3] Brügger, N. 2012. Web History and the Web as a Historical Source. In: ZeithistorischeForschungen / Studies in Contemporary History, Online Ausgabe, 9 (2012), H. 2: http://www.zeithistorische-forschungen.de/16126041-Bruegger-2-2012.

[4] van den Heuvel, Ch. 2010. Web Archiving in Research and Historical Global Collaboratories. In: Brügger, Niels (ed.): Web History. New York 2010, p. 279-303.

[5] Hockx-Yu, H. 2013. Scholarly Use of Web Archives: http://files.dnb.de/nestor/veranstaltungen/2013-02-27-scholarly-use-of-web-archives_public.pdf.

[6] Meyer, E., Thomas, A. & Schroeder, R. 2011. Web Archives: The Future(s): http://netpreserve.org/resources/web-archives-futures

[7] Meyer, E. 2010. Researcher Engagement with Web Archives: Challenges and Opportunities: http://repository.jisc.ac.uk/543/1/JISC%2DREWA_ChallengesandOpportunities_August2010.pdf

[8] Cremer, M. 2013. Providing Access to the DNB Web Archive: http://files.dnb.de/nestor/veranstaltungen/2013-02-27-providing-access-to-the-DNB-web-archive.pdf

[9] http://webis.sub.uni-hamburg.de/webis/index.php/Wissenschaftliche_Bibliotheken

[10] http://www.dfg.de/download/pdf/dfg_im_profil/evaluation_statistik/programm_evaluation/studie_evaluierung_sondersammelgebiete_empfehlungen.pdf

[11] Spaniol, M., Mazeika, A., Denev, D. & Weikum, G. 2009. Catch me if you can: Visual Analysis of Coherence Defects in Web Archiving. In: The 9th International Web Archiving Workshop (IWAW 2009). Workshop Proceedings, p. 27-38: http://www.iwaw.net/09/IWAW2009.pdf

[12] Gray, G. & Scott, M. 2013. Choosing a Sustainable Web Archiving Method: A Comparison of Capture Quality. In: D-Lib Magazine 19, Number 5/6 (2013): http://dx.doi.org/10.1045/may2013-gray

[13] Niu, J. 2012. Functionalities of Web Archives. In: D-Lib Magazine 18, Number 3/4 (2012): http://dx.doi.org/10.1045/march2012-niu2

[14] Pennock, M. 2013. Web-Archiving: DPC Technology Watch Report 13-01 March 2013, p. 23: http://dx.doi.org/10.7207/twr13-01

[15] http://www.webarchive.org.uk/ukwa/info/mementos