News Media Section

Selective Web Archiving at the German National Library

Tobias Steinke
Informationsinfrastruktur, Deutsche Nationalbibliothek, Frankfurt am Main, Germany.
t.steinke@dnb.de

Abstract:

The German National Library has been collecting web sites since 2012 on a regular basis together with a service provider. The workflow includes selection of web sites, addition of metadata about title and categories, crawling, integration in the catalogue, giving access in the reading rooms and archiving. Web sites of selected institutions like federal authorities, interest groups and cultural institutions are normally crawled twice a year. There are also event based crawls about e.g. elections, sports and cultural events. In 2013 the complete online offering of the Financial Times Deutschland was captured before it disappeared as a result of the closedown. These experiences helped getting a better understanding of the challenges of archiving news sites.

Keywords: web archiving, selective crawls, german national library, news sites.

Web Archiving at the German National Library

In 2006 the federal law about the national library in Germany was revisited and enhanced. Since then there is a legal deposit in Germany including non-physical media works (online publications)[1], which are collected, catalogued, indexed and archived.

Digital publications were already collected before 2006 on data carriers like floppy disks, CD-ROM's and DVD-ROM's. This includes multimedia publications, audio CD's and educational software, but no games. After the enhancement of the law the collection includes also e-journals, e-thesis, e-papers, e-books, digitized books, digitized music and music files. All these publications are directly delivered by the publishers to the German National Library

[1] Gesetz über die Deutsche Nationalbibliothek: http://www.gesetze-im-internet.de/dnbg/index.html

(DNB) via several technical interfaces[2]. Access to the publications is given at least in the reading rooms in Frankfurt and Leipzig. If the right holder granted it, access is also possible on the web.

The collection of web pages is a new challenge. A direct delivery from the publisher to the library is not feasible because web pages are often dynamical created out of content management systems and highly interlinked with other publications. Web pages are also never in a finished state, the content may change often in parts or completely. To deal with this challenge it is common to collect web pages with a software tool called crawler or harvester. A crawler visits a web page, saves the content, follows all links on the page to other pages and visits these pages until an end condition is reached (e. g. links will no longer go to the same domain). The crawl of a certain page is repeated frequently (e. g. two or four times a year) to capture the ongoing changes. Nevertheless the results are just snapshots of the web pages, which might miss temporary changes in between the visits. The technical concept of crawlers does also not work for very dynamic content, especially content that is created as a result of user input. So the results of web harvesting are neither comparable to common library collections in regard to quality nor to completeness.

Many national libraries do web harvesting and several of them are in the International Internet Preservation Consortium (IIPC)[3]. 50 libraries, archives, academic institutions and companies share their experiences on web archiving in this organization. At conferences and in working groups tools and concepts are developed and discussed. The German National Library has been member of the IIPC since 2007. This membership helped understanding and preparing a workflow for web archiving to deal with the new task. Although DNB did two event crawls in 2005 and 2007 with the European Archive (now called Internet Memory Research), there was no regular workflow for web archiving.

There are two common strategies in web archiving: Top-level domain crawls and selective crawls. As the web is international it is very difficult to collect just a national part of the web. The legal deposit is about German publications. But there are seldom imprints on web pages, so it is not easy to identify publications from Germany. One way of getting at least many web pages which were probably published in a specific country is to collect web pages with URL's containing the assigned national domain (e. g. .fr for France or .de for Germany). But it is possible for a German publisher to register a URL for the web site with a different top-level domain and for a French publisher to register a URL with .de. The top-level domain is just one possible indication of a web site of a certain country. Many national libraries did and do domain crawls of their national top-level domains on a regular basis. The number of registered domains for each top-level domain varies significantly. E. g. there are currently 56 thousand registered domains for .is (Iceland), 2 million for .fr (France) and 16 million for .de (Germany). The effort and the needed resources for a top-level domain crawl vary accordingly. Therefore DNB decided not to start with a top-level domain crawl. It was done eventually in 2014 by the Internet Memory Research[4] for DNB as supplement to the selective crawls[5].

[2] Description of the submission options: http://www.dnb.de/EN/Netzpublikationen/Ablieferung/ablieferung_node.html

[3] http://netpreserve.org/

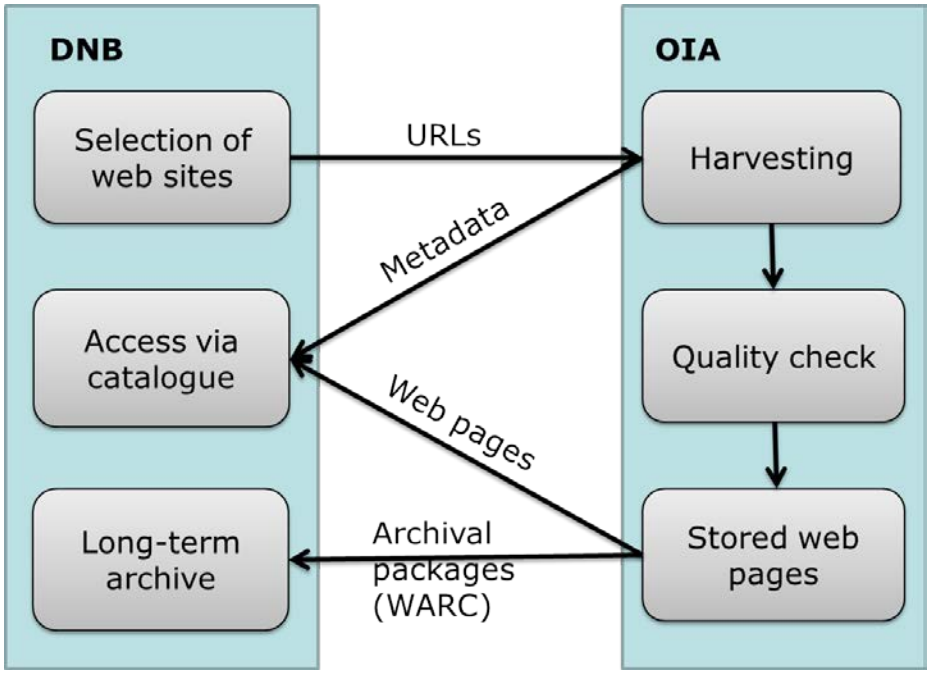
[4] https://internetmemory.net/en/

[5] 6 million sites were collected with 2.4 billion files and 120 TB of data.

Selective crawls are focused crawls of certain web sites (all pages of a certain domain like www.dnb.de) or web pages (single pages). The selection could be according to topics or events. The selection process is manual and often involves finding the best initial parameters for the crawler to get the intended results. The resulting snapshots could be handled similar to periodicals. Selected web sites and pages are collected in many libraries. In 2010 DNB started with an internal project to compare the options for every step in a workflow for selective web archiving. As a result it was decided to outsource most of the steps. In a call for tender in 2011 the German company oia[6] was picked as a service provider.

Workflow for Selective Web Archiving

A workflow of a web archive includes all the usual steps in a library workflow: Collection, indexing, archiving and making it available. For selective web archiving this means selection, harvesting, quality assurance, storing, indexing, preserving and giving access. DNB built up such a workflow with the service provider oia.



Selection: Librarians at the German National Library select web sites according to topics or events. They use a tool by oia on their local work stations to determine the starting point for a crawl (a URL). It is also possible to adjust the default crawler parameters for each site (get everything from the same web site, repeat crawl twice a year). A title is manually given for every site because there is no reliable way of automatically getting the title of a web page. One or more categories are manually assigned to each site. A category could be a topic like “federal authority” or an event like “Bundestag elections 2013”.

Harvesting: A crawl task is submitted from the tool installed at DNB (in Frankfurt and Leipzig) to oia (in Düsseldorf). The crawl is scheduled by oia according to selected requirements and resources balancing. The actual harvest is done with a self-developed

[6] http://www.oia-owa.de

crawler by oia on servers of oia. The crawler identifies itself as working on behalf of DNB and respects the robots.txt mechanism.

Quality assurance: oia checks manually the basic quality of the crawl results with a best effort approach. The crawl is repeated if needed.

Storing: The checked crawl results are stored on servers of oia. The storage is paid by DNB and optimized for giving access.

Indexing: The input data for each crawl (URL, title, categories) together with the crawl data and access ID are transferred back to DNB from oia via an OAI interface. The received metadata creates automatically a new entry in the library catalogue. There is a separate catalogue entry for each web site that lists all crawls of it. oia creates and updates also a full text index of all web pages in the selective web archive.

Preserving: All crawled data is provided by oia as WARC files. WARC[7] is an ISO standard of an archival container format for web archiving. The WARC files are transferred to DNB for storage in the digital preservation system.

Giving access: All archived web sites are exclusively accessible in the reading rooms of DNB. Access on the web would be possible with permissions of every right holder, but it is not feasible for DNB to get all of these. Users in the reading rooms can access the crawls via the catalogue entries or via a separate portal page provided by oia. This portal page offers a browsing interface based on the categories and a full text search. The full text search enables restrictions on crawl date intervals, categories and sites.

The regular crawls started in 2012 with a small selection of web sites. Topic related categories are mainly federal, cultural, research, sports and religious institutions. These sites are crawled twice a year. Sites related to events are just crawled a few times around the time of the event. This includes political, sports, cultural and disaster events. There are currently 1,700 web sites with 8,300 crawls in the web archive. It is planned to work together with existing collections of topic related sites to expand the range of collected sites.

Crawling of News Sites

News sites are a specific challenge. They are updated very frequently, often several times in an hour. New articles are constantly added and links to existing ones disappear from the start page. Even if the article is still there, it will not be harvested by a crawler without a link. Several news pages have so called pay-walls. These are restrictions on access to some or all of the pages. A login is needed to access the restricted parts of the site. There are many different restriction models. A page could be freely accessible for a few days before it gets restricted. Sometimes an article is available in a free shorter version and in a restricted longer version, but both versions have the same URL. Restricted pages and especially the variations of restrictions make it very difficult to crawl.

In 2013 the German version of the newspaper Financial Times (Financial Times Deutschland) was closed. Their news site www.ftd.de existed a few weeks longer before it vanished. DNB decided to harvest the complete web site in these weeks. That was the first

[7] http://www.iso.org/iso/catalogue_detail.htm?csnumber=44717

experience with a crawl of a news page. It was not as challenging as other news sites because there was no more updating and there was no pay-wall. But there were many articles which were no longer linked from the start page. To address this DNB contacted the site holder and they were able to provide a list of links to all pages. Then oia modified their tools to be able to handle a huge list of direct links as a starting point. With this modification they managed to capture all the content of the site right in time. Full access to all collected articles is restricted to the full text search by oia, because the link in the catalogue leads to the (last) start page without the links to all the old articles. The experiences of this crawl were very helpful for the on-going discussion on expanding the collection with news sites.

In 2014 DNB organized a workshop with representatives of the biggest German news sites. All agreed on the interesting observation that they never delete articles. The old articles are normally available as long as the news site exists and are accessible permanently via the same links. So if these direct links are known and could be used for crawling, a complete crawl of the site should be possible any time. Some of the publishers pointed to Google sitemaps[8] as a way to get the needed list of links. A Google sitemap is an XML file with a list of links which a site holder wants Google to crawl. But this file could be submitted directly to Google, so it is not necessarily publicly available. Nevertheless most of the publishers seem to be able to provide a Google sitemap. There was no helpful outcome on the topic of pay-walls. Some of the publishers were very skeptical on the idea of letting crawlers behind their pay-walls. It was confirmed that the concepts of access restrictions differ and a straight forward crawling would not work.

To test the crawling of a still active news site there was recently a test crawl of one of the biggest German news site, SPIEGEL ONLINE[9]. There is a Google sitemap for this site and this could be used as an input for the tool by oia. The first crawl did not give the expected results because the default crawling parameters did not work quite well for the huge amount of links. A modified crawl will be done shortly.

But even if the best configuration for these kinds of crawls could be found, there are still a lot of open questions to discuss. How often does it make sense to do a full crawl of a news site? Should we do a crawl of just the starting page very often? Is it enough to get access to most of the articles just by full text search? And how do we handle restricted access to many web pages especially on news sites? It is probably needed to find individual solutions in cooperation with the publishers.

The selective web archiving workflow of the German National Library turned out to work well for many web sites. In the upcoming years the collection will be expanded and news sites will become an important part of it.

[8] https://support.google.com/webmasters/answer/156184?hl=en
[9] http://www.spiegel.de/