Archiving the Czech Web: Issues and Challenges

Petr Žabička[1]

[1] Moravian Library in Brno, Kounicova 65a
601 87 Brno, Czech Republic
zabak@mzk.cz
http://www.webarchiv.cz

Abstract. The Czech web archiving project started as a two-year research and development project of the National Library. Its main target was to investigate problems connected with collection and long-term preservation of electronic information resources. The project team was able to establish a testbed for harvesting web resources, based on software tools developed by the NEDLIB project. Nowadays, the Webarchiv project continues as a collaborative effort of the National Library of the Czech Republic, Moravian Library in Brno and Masaryk University. This paper describes the project and discusses its possible future development.

Introduction

In the Czech republic, as in many other countries, the National library is the institution entrusted with the task of protecting and conserving written national cultural heritage. The growing number of digital born documents therefore could not escape its attention. Unfortunately, Czech legal deposit law did not take digital documents into consideration when it was formed in the nineties. Aware of the growing danger of losing a large portion of the national digital heritage the National Library initiated a two-year research and development project "Registration, protection and access to national electronic resources on the Internet", funded by the Ministry of Culture.

Research Project (2000-2001)

The aim of the project was to expand the scope of the Czech National Bibliography by adding electronic documents, concentrating mainly on online documents. It was clear from the beginning that it will be necessary not only to identify documents worth accepting into the National Bibliography. It was necessary to create a permanent repository for online electronic resources.

The project team was aware of many obstacles on the way: The large number of electronic documents called for new approach to their bibliographic description. The technological dependence of online documents called for know-how that was in short supply in an institution dealing mainly with traditional printed material. To counter

these problems, the National Library joined forces with the Institute of Computer Science of Masaryk University in Brno, which maintains Czech version of Dublin Core specification and is also active in digital library research.

After thorough analysis of ongoing projects dealing with online documents contacts were made with Helsinki University Library. Building on their experience, we decided our priority was to start breadth-first harvesting of the national web space. This decision was based mainly on our need to preserve as large portion of our national digital heritage as possible. To this end, we acquired a linux server and installed NEDLIB Harvester software [1]. Unfortunately, the project funding proved insufficient to provide for large storage capacity needed for the document repository. The only mass storage system available to us was a tape robot based archive holding images from National Library's digitization projects. We knew that the tape robot will have to be replaced by another kind of mass storage device to allow real-time access to the archive but we were sure it is only a matter of time before such a device with sufficient capacity will become affordable.

While the web harvesting infrastructure was being prepared, the rest of the team worked on the bibliographic track of the project. After analyses and experiments with Dublin Core based cataloguing of web resources we prepared set of web archiving criteria:

- Top-level domain based. Useful as a main limit for breadth-first harvesting, this rules out any domain outside the .cz domain.
- Language based. Web pages in Czech will be accepted if they are recognized as such either automatically or manually. This rule cannot be used for breadth first harvesting yet as the automatic language recognition has not been implemented into the harvester.
- Content based. For focused harvesting only scholarly, scientific or artistic documents will be accepted as we expect wider audience for this kind of documents. We disregard private web pages, company presentations and advertisements.
- Bibliographic category based. We prefer serials, conference papers, research papers, scholarly works and other gray literature.
- Form of publication based. We accept resources published in electronic form only to prevent duplicate handling of paper and printed versions of the same document.
- Access based. We accept only publicly and freely available resources.
- Format based. We prefer widely accepted documents formats like html, xml or jpg and de facto standards like pdf or MS Office formats.

To promote the use of metadata in online published content, we implemented, localized and further developed several tools such as Dublin Core Metadata Generator or URN:NBN unique identifier generator.

In September 2001 first pilot crawl of the .cz domain was attempted. Although the crawling had to be stopped after two months due to software problems we have managed to collect nearly 130 GB of documents. Experience with the first crawl led us to a decision to develop Harvester Configurator – web based interface to the harvester. Main motive for this decision was to allow librarians themselves to set up and control harvester crawls and, if necessary, to run several parallel crawls with different parameters. To facilitate this we had to modify the harvester to allow for more detailed configuration without the need to recompile of the whole package every time a parameter has been changed.

At this stage we encountered first problems related to the size of the harvested archive: the harvester created database containing information about the current crawl grew so large that it was possible to process only the most simple queries in reasonable time. As the documents themselves were stored on the slow tape robot, it became increasingly difficult to keep track of the archive contents. We were also aware of the prohibitive cost of commercial indexing engines like the one used in the Nordic Web Archive [3] project. We were not sure the project would get any substantial funding so we had to find a solution that would carry minimal additional costs.

Before the end of the year we asked a group of students from the Charles University in Prague to develop an open source indexing engine and search, retrieve and present interface for the web archive. Although the students had to complete this task as part of their course of study, we knew it will take more than a year to finish and that the results might not meet our expectations. On the other hand we felt that their success will give us the key application, which was missing from the project infrastructure.

At the end of 2001 the research project was concluded and the grant commission approved it.

WebArchiv Project (2002)

Although a follow-up project has been prepared during 2001, there was no funding for new R&D projects available for 2002. The work of the project team was limited to a minimum necessary to keep the infrastructure working. New project was therefore prepared and application was submitted into a one-year library development program. Funding under this program was awarded to WebArchiv in the spring.

The harvester development did not stop though and in April 2002 second harvest of the .cz domain has been initiated. This time we decided to cover the domain as completely as possible. That meant going into a greater depth and accepting URLs with parameters. The harvesting started without problems but soon we saw that system throughput is limited – two months into the harvest we were able to collect only about 5,5 GB of data each day over a 100Mbps internet connection.

By July, we collected about 10 million documents (240GB of data) from about 30,000 second level domains. This number is quite impressive but there were about 130,000 registered second level domains under the .cz domain at that time. We assumed many of them were just blocked by domain speculators but to make sure that we did not miss anything important we obtained a complete listing of registered second level .cz domains by querying of the major name servers for .cz. (Later experiments have shown that out of 140.000 registered domains 120.000 have web server running on http://www.*domain*.cz and 76.000 have web server running also on http://*domain*.cz.).

Our harvesting experiments have shown several weaknesses of the harvester. One of the most profound was the inability to detect long lasting local network outages – harvester should freeze harvesting automatically when local network or Internet connection fails. We are also thinking about shortening pauses between attempts to download documents from one server when there are just few servers to harvest from

during focused crawls. We also miss the possibility to set different harvesting depths for individual servers and support for javascript. On the other hand we were able to boost the harvesting performance by installing DNS cache. Although the harvesting can be speeded up even further by giving list of all .cz domains with running web server to the harvester, we are not sure everyone will welcome this as many of the newly registered servers might not be ready to be published at the time of harvesting.

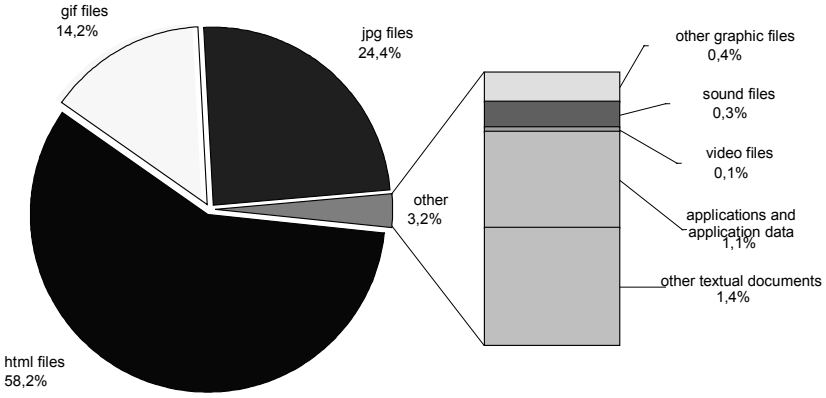

Fig. 1. Relative distribution of archived files by file type

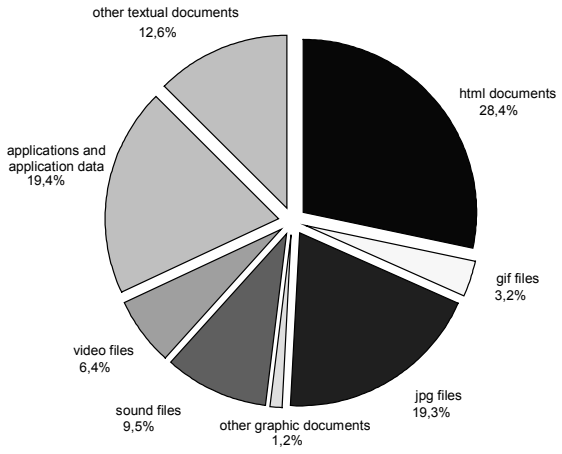

Fig. 2. Absolute distribution of archived files by file type

In the autumn several harvester bug fixes were released. One of them fixed a problem of incorrect calculation of MD5 checksum for archived documents. Check of

the archive proved that indeed some of the checksums in the archive were incorrect. To secure archive consistency it was necessary to reread the whole archive and check its consistency. During this process we found more inconsistencies than we expected and thanks to the tape robot slowness it took us several weeks to repair the whole archive.

The need for speed and reliability was one of the main factors behind our decision to look for new ways of data storage. We started talks with Masaryk University team involved in grid research with the aim of using datagrid storage capacity for the web archive. Although they could not yet give us the level of reliability we needed they offered us a large portion of their disk array so we could transfer the archive off the tape robot. By the end of the year new server was up and running in Brno, plugged straight into a node of the CESNET academic Internet backbone.

Integrated Project (2003) and Future Work

The National Library, together with the Charles University's Institute of Computer Science, constitutes the core team that operates Uniform Information Gateway – library information portal for Czech libraries, based on Metalib and SFX products by ExLibris. For the coming year, WebArchiv integration into National Library processes and namely the Uniform Information Gateway is among the priorities. WebArchiv project development was therefore incorporated into an integrated subject gateway project of the National Library and submitted under a one-year library development program.

Because the position of the National Library regarding legal deposit of online content is not strong, the National Library started talks with selected publishers at the beginning of the year. The aim of the talks is to persuade the publishers to cooperate with the National Library by publishing Dublin Core or other metadata in their online content and to let the National Library not only harvest their web sites but also to make the archived documents available to library users. This initiative met with moderate success with about 15 agreements signed and we expect more during the second half of this year. Some of the publishers even agreed to add metadata to their web content and many of them wondered why do they have to sign contract that limits the public access to the archive to a dedicated workstation within the library. On the other hand there were publishers who refused to sing any contract with the national library at all. Unfortunately, agreements with the publishers are now the only way of giving access to at least a portion of the archive. Although the access will be limited to local users of the National Library, it will give us valuable feedback for future work.

In June, after 15 months of work, the Charles University students have completed the browse and search interface and the underlying indexing engine for the archive. Their package indexes only html files but its API is open for other format modules. It supports Czech grammar specifics, it is aware of diacritics and character translation issues. It supports 9 search categories including fulltext, emphasized text and metadata. Unfortunately, this software does not support any of the major library standards as the software development proved to be challenging even without them. That leaves us with the necessity to add support for Z39.50, OAI-PMH or other standard our-

selves as it is very likely the authors will not be interested in further support and development of their product.

Cooperation

To embed Czech web archive tightly into the Czech library infrastructure another step will be taken: In cooperation with Czech ISSN agency we will create bibliographic database of Czech online periodicals. This database will contain both periodicals registered by the ISSN agency and periodicals registered by the WebArchiv project as important sources for the National Bibliography. The WebArchiv and the ISSN agency will also cooperate by informing the publishers they communicate with about the existence of the other partner.

The WebArchiv project started in 2000, which was also the closing year of the NEDLIB project. Although we had contact with some of the NEDLIB members and gradually we became aware of more European web archiving initiatives, we feel the need for closer cooperation.

As the National Library could not afford the price of cooperation with the Internet Archive [4] and as it is directly involved in several European projects already. Therefore, a new partner joined officially the WebArchiv project – the Moravian Library in Brno. Both institutions agreed that the Moravian Library will become the official WebArchiv representative for international cooperation and that it will also be responsible for the technological side of the project while the National Library will represent the project on the national level. Both libraries will continue their cooperation with the Masaryk University, which provides much of the programming skill for the WebArchiv.

Conclusions

We have presented the Czech Web Archive project, its development and plans for the future. We have provided overview of the project infrastructure developed so far. It is clear now that without legal deposit law for online publications only limited funds can be spent on making the archive publicly available.

We have also shown the file type distribution on the Czech web that corresponds with findings of other researchers. It is very likely that by further analysis we can find criteria that will limit the size of the archive substantially while not diminishing the historical value of the collection.

We have shown that it is possible to start archiving the web with little funding but that any activity has to have clear support on the part of the national legal deposit authority. In the Czech Republic, the National Library plays the leading role in the web archiving activity. Unfortunately, other types of memory institutions, namely archives and museums are not yet interested in collecting the digital material. In their absence, only cooperation with other libraries and universities can keep the project running.

References

1. J. Hakala. Collecting and preserving the web: Developing and testing the NEDLIB harvester. RLG DigiNews, 5(2), April 15 2001.
2. NEDLIB. Website. http://www.kb.nl/coop/nedlib/.
3. Nordic Web Archive. Website. http://nwa.nb.no.
4. The Internet Archive. Website. http://www.archive.org.
5. WebArchiv. Website. http://www.webarchiv.cz.