Archiving the Slovenian Web: Recent Experiences

Alenka Kavčič-Čolič [1], Marko Grobelnik[2]

[1] National and University Library, Ljubljana, Slovenia
alenka.kavcic@nuk.uni-lj.si
http://www.nuk.uni-lj.si/
[2] Jozef Stefan Institute, Ljubljana, Slovenia
marko.grobelnik@ijs.si
http://www.ijs.si/

Abstract. The National and University Library and the Jozef Stefan Institute are collaborating in a joint project, which is aimed at the development of a national repository for long-term preservation of Slovenian web and electronic resources by Slovenian authors. This two-year project, which is to be completed in September 2004, was sponsored by the Slovenian government. The project has two main goals i.e. to develop a methodology for archiving the Slovenian web and to build a web crawler. In this paper we describe the recent experiences.

1 Introduction

The National and University Library (NUK) has the function of the Slovenian national library and as such it has the mission to collect, protect and preserve Slovenian written heritage on all carriers. The documents published on the web and other electronic sources are part of this heritage and NUK has the duty to preserve them for future generations. Due to the lack of financial resources and computer science expertise, the implementation of this task had to be postponed for several years.

In October 2002 NUK applied for financial support in a tender to finance research projects, organized by the Ministry of Culture, Ministry of Education, Science and Sport and the Ministry of the Information Society. The aim of NUK's project was to develop a web archive with a complete archiving methodology. At the same tender the Jozef Stefan Institute (IJS) presented its candidature as well. IJS is the main research institute in the field of natural sciences, physics, computer sciences and artificial intelligence in Slovenia. Their project consisted in the development of a web crawler. We were informed about our common interests by the ministries and with their mediation we joined our research teams and prepared a joint project which consisted of two parts: the research team of NUK had the task of developing the library aspect of the project i.e. the processes of selection, cataloguing and archiving the web, while IJS focused on the computer science aspect, that is the software development of the tool for crawling the web.

In continuation both parts are briefly described.

2 Methodological Approach to Slovenian Web Archiving

NUK's research team aimed to achieve two different goals:

- design of a proper model for archiving electronic sources according to their different types; and
- preparation of specifications, rules and standards for the development of an electronic repository as well as for the processing, collection, cataloguing and long-term preservation of electronic documents, i.e.:
  - definition of the selection criteria for electronic sources
  - preparation of protocols for copyright management (negotiation and agreements)
  - specification of the bibliographic database
  - definition of software and hardware requirements
  - definition of location and infrastructure requirements
  - definition of processes and human resources needed
  - definition of the practice for long-term preservation.

2.1 Harvesting methodology

At the beginning of the project we were faced with a dilemma: the Slovenian web should be preserved, since it represent a cultural, sociological, historical and technological document of the nation. Besides, we cannot anticipate what will be of interest to our users in the future. On the other hand, there are a lot of electronic sources published on the web that are very important to today's users and it is necessary to deal with them separately. Among these there are contributions to conferences, online master's degree theses, teaching materials, legislation, public sector documents, newsletters, etc. Many of them are published on the internet only. The experiences with our users show that if we do not catalogue these sources and keep them separately, they would hardly find them even if they are available on a portion of the archived web. Moreover, a great many of the electronic sources contain dynamic pages or the access to them is limited and therefore they cannot be captured by the crawlers.

The review of the literature of various web archiving experiences such as NEDLIB [16], Kulturarw3 [10], Combine [6], Nordic Web Archive [17][3], the experiences of the French National Library [1] and Czech Republic [24] in Europe, Internet Archive [9] and MINERVA [13] in the United States, the Greenstone Digital Library [7] in New Zealand and PANDORA [19] in Australia has shown us that there is no best solution and every one of the mentioned experiences have their advantages and disadvantages. The French have shown that it is possible to combine two opposite strategies of archiving: the selective (focused) and non selective approach. This solution has attracted our attention and we have been looking to implement a similar solution to our archive.

The Slovenian public web is very small, it is expected that a crawling by domain will take no more than 1 Tb. The access could be solved with an additional module for indexing, searching and browsing. The problem is that many valuable electronic sources would not be recognizable or localizable and their authenticity would be therefore doubtful. Consequently it was necessary to add the possibility of collecting the electronic sources separately, specially those with limited access, dynamic pages, or other functionalities.

For this purpose it was necessary to built a repository for electronic documents consisting of a document management system and modules for delivery, long-term archiving, metadata searching and access in addition to the web archive. In this context we designed a model of the electronic repository according to the standard ISO 14721:2002 or reference model OAIS [18]. The experiences in the project NEDLIB were very helpful in this regard.

2.2 Selection criteria

Another issue that we had to solve out was the definition of the selection criteria. For capturing the web pages the selection was limited to the *.si domain and other domains such as *.com or *.org with Slovenian language identification techniques. Major task was to specify the selection criteria of the online electronic sources for the repository. At the beginning we decided to extend the praxis of cataloguing to the electronic media. NUK together with 270 other Slovenian libraries is cooperating in a shared cataloguing system, named COBISS, which uses the COMARC format, a Slovenian adaptation of UNIMARC. In the COBIB (one of the COBISS system databases) there are approximately 2.4 million of bibliographic records, covering monographs, serials, journal articles, maps, audio, video and other materials. At the beginning it was decided to focus on a few publications types like electronic monographs, serials, maps and web pages. The definition of the typology has not been completed due to the introduction of the concept of integrating resources to which belong the web pages.

2.3 Identifiers

For some years now the electronic sources are catalogued in the union shared cataloguing system in UNIMARC, producing an URL link to the document in the 856 field. It would be more secure if the bibliographic record would be linked to the document kept in a repository. For this purpose we decided to use existing bibliographic identifiers, such as Uniform Resource Names (URN) according to RFC 2288 [11]. In the case of journal articles we plan a combined URN with source data. The automatically harvested web pages and other electronic sources without bibliographic identifiers would obtain an URN containing a checksum. It is still unclear whether to include national bibliographic numbers to the electronic sources or not, since the national bibliography is selective and does not cover all Slovenian production.

2.3 Metadata

The web pages and other documents in the web archive will be indexed very similarly to those, indexed by various browsers on the internet. In the repository the electronic sources will be kept separately from their metadata. It will speed up the indexing and metadata search. We decided to use UNIMARC format for the bibliographic database structure. It has been simplified as much as the main fields were retained to allow for data export to the union catalog and other local catalogues, such as the national bibliography database. Although UNIMARC is a very complex format it contains the 856 field where we can describe the characteristics of the electronic source. Upon a comparative analysis of different experiences (CEDARS [4][12], OCLC/RLG Working Group on Preservation Metadata [23], NEDLIB [15], National Library of Australia [21][22]) we added other fields that are important for long-term preservation. A conversion to Dublin Core has been planned as well in order to support interchange with other systems.

2.3 Copyright management

Archiving web pages requires attention on three aspects which have different legal basis: (1) the procedure of crawling the web pages and electronic documents on the internet; (2) enabling the public to access them; and (3) their preservation for the future. Since the Slovenian Legal Deposit Law does not cover intangible electronic publications the harvesting process should be regulated by the copyright legislation.

NUK took part in the preparation of the new Legal Deposit Law which includes publications on all carriers. We cannot foresee when this law will pass in Slovenian Parliament since the publishers' lobbies are still very strong. In the meanwhile we have to implement the Copyright Law to the three processes of archiving. In this regard we disseminated the Code of practice for the voluntary deposit of electronic publications [5] among Slovenian publishers and tried to find common interests with them in the deposit of electronic publications. We also prepared sample agreements for different scenarios and discussed it with a few publishers.

2.3 Repository infrastructure

At present there are more than 20 servers in NUK. Since there is no one among them with the capacity for storing such volume of data, it has been necessary to purchase a stronger server Compaq NAS B3000 with 1Tb of disk-space with a possibility of extension. A back-up system is included. The locations and furniture necessary for back-up archiving and safety procedures are being planned as well as the possible strategies for long-term preservation.

3 Crawling the Web

Crawling the web is fundamental technique for collecting the data from the web. In general, we distinguish several types of crawlers, most significant groups being: (1) crawlers which try to collect as much as possible of the data from the web, and (2) focused crawling which collects just specific parts of the web (e.g. based on the pre-selected topic or specific language). There are a lot of important issues related to the crawling, most important being connected to the indeterminacy of the web. Indeterminacy could be seen from (1) technical point of view, as the data is constantly growing and changing and the main problem is what and how to collect, and (2) from librarian point of view, on how to maintain such a collection through appropriate meta-data etc.

In our case, the goal is to archive everything what is relevant for the Slovenian which is still within reasonable cost. Such a practical definition of “Slovenian” web which we used in building the system and collection was the following:

(1) Slovenian web is everything what is physically located in Slovenia,
(1) everything what is written in Slovenian language and resides anywhere,
(1) all the documents which contain terms relevant for Slovenia.

First two items can be easily covered by using classical crawling techniques in addition to language identification techniques (for the second item). The third category of documents could be covered by inquiring one or several of major global search engines using focused crawling techniques. In the first phases of the project we decided to cover first two items since they cover most of the relevant materials for archival.

Depending on the specifics of the definition what exactly to crawl from the web, we estimate (based on the experiments) the size of Slovenian web is approx. 10 millions textual pages which makes it together with images and other file types (we store all the accessible data) in the range of 100Gb up to 1Tb.

One of the fundamental decisions, when crawling the web, is the strategy of accessing (URL ordering) the web pages. In our case we use breath-first [14] search with prioritizing URLs with PageRank [20] estimate (used in Google) which gives good results. The effect of this strategy is to prioritize more important pages.

3.1. Data Storage

For the storage of the crawled pages we used similar philosophy as in Google [2]. We avoided storing the data into standard relational databases with relatively big and unnecessary overhead. We rather built our own small and compact storage engine which allows efficient basic operations on storing deleting and accessing the data on the single and distributed platforms.

3.2. Storing and Accessing Multiple Versions of the Same Page

The purpose of crawling the Web for archiving is to be able to reconstruct the contents of the web for some time-point in the past. Therefore, we need to be able to store in the database multiple versions of the same page as well as being able to retrieve the contents for specific time-point. Each Web page is annotated with information about the time of crawling which is further used when accessing the page. We use this information when addressing the page through redirection URLs. The pages in the data-base are stored in the original format as seen at the time of crawling and are converted on the fly when transferred to user's client with URLs being replaced by redirection URLs including information about their crawling-time – this enables accessing all other elements on the page from the same time-point as the target page.

3.3. Browsing and other Analytic Tools

Once the data is being collected from the web there is an immediate next issue what and how to deal with it. There are standard ways what one can do with the data, such as adding meta-data according to librarian schemas, full text indexing for the purpose of the search etc. In our project we would like to go beyond classical techniques using modern data processing and data analysis techniques based on Text-Mining, Link Analysis, Semantic Web, Information Retrieval and other related research areas. The main characteristic of such techniques is to approach the data from various angles using all kinds of information within the data such as text semantics, linkage (hyperlinks), linguistics, temporal data, etc. In the following sections we will present some of the approaches we plan to implement in the course of the project.

3.4. Taxonomy Construction

One of the more popular ways how to organize the contents of the data is hierarchy of topics. On the web, there are several popular taxonomies such as Yahoo (http://www.yahoo.com), Open Directory/DMoz (http://www.dmoz.org), Medline/MESH (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi), etc. These and other similar taxonomies organize the documents (usually manually) into a tree of topics available for manual browsing. Recently in the areas of Text Mining and Semantic Web appeared as an important research topic so called "ontology construction" having as one of the problems also (semi)automatic construction of taxonomy tree. We will invest some of the resources into adapting this kind of techniques for organizing pages into a tree of topics in a semiautomatic manner with the help of machine learning techniques "Active Learning" and "Labeling Unlabeled Data".

3.5. Text Categorization

A scenario which is even more useful is many practical situations is how to categorize a textual document into a predefined set of categories. Categories could be organized hierarchically or flat. As a follow-up of a recent work of the project collaborators we'll adapt a system built for categorizing documents into Yahoo taxonomy [8]. The idea is to automatically maintain hierarchically organized index of topics and for most of the crawled pages having precalculated categories from the index.

3.6. Information Extraction

Information Extraction is a research subfield of Text Mining having relatively large commercial success because of practical use of its results. The goal is to extract from the documents pieces of information which are the most relevant for understanding of the document contents. The most popular category of entities being extracted are so called "Named Entities" representing names of people, companies, geographical names etc. Having a document represented with named entities instead of the full text enables us to use other analytic techniques for discovering the contents in the document corpora. Here is an example from Slovenian daily newspaper "Dnevnik" from Saturday 31.3.2001 describing president Milosevic being put on the court:

Primer ***Milo_evi_***
Vladni viri: ***Milo_evi_*** *v sodni pala_i; zasli_anje jutri? Neimenovani srbski vladni viri so za* ***Radio B92*** *potrdili, da so* ***Slobodana Milo_evi_a*** *_e prepeljali v beograjsko sodno pala_o, in pojasnili, da je bil nekdanji jugoslovanski predsednik aretiran "ne le zaradi nezakonite gradnje, ampak tudi zaradi razli_nih finan_nih nepravilnosti". Po poro_anju srbske dr_avne televizije* ***RTS*** *naj bi* ***Milo_evi_a*** *zasli_ali jutri. Uradno srbska vlada vesti o* ***Milo_evi_evi*** *aretaciji sicer _e ni objavila, novico pa je za* ***Radio B92*** *potrdil podpredsednik* ***Socialdemokratske unije****, ene od strank vladajo_e koalicije* ***DOS****,* ***Vlatko Sekulovi_****.Malo pred polno_jo so v sodno pala_o pripeljala _tiri vozila, v enem od njih naj bi pripeljali* ***Milo_evi_a****. Po poro_anju srbske tiskovne agencije* ***Tanjug*** *so omenjena vozila pred tem videli pred* ***Milo_evi_evo*** *vilo na* ***Dedinjah****. Jugoslovanskega predsednika* ***Vojislava Ko_tunice*** *trenutno ni v dr_avi, saj se mudi na konferenci o razoro_evanju v* ***_enevi****.*

Named entities being extracted from the articles are boldfaced above and listed below. Each named entity has attached frequency and surface forms as appeared in the documents.

Slobodan Milo_evi_ (7): Milo_evi_, Slobodana Milo_evi_a, Milo_evi_a, Milo_evi_evi, Milo_evi_evo
Radio B92 (2)
RTS (1)
DOS (1)
Vlatko Sekulovi_ (1)
Socialdemokratska unija (1)
Dedinje (1)
Vojislav Ko_tunica
_enevi (1)

Taking this kind of information from many articles (in our case 50.000) and put it into one picture we can get a graph how named entities are related to each other. In Figure 1. we see Slobodan Milo_evi_ being put in the middle surrounded with the most frequent named entities appearing with him in the context.

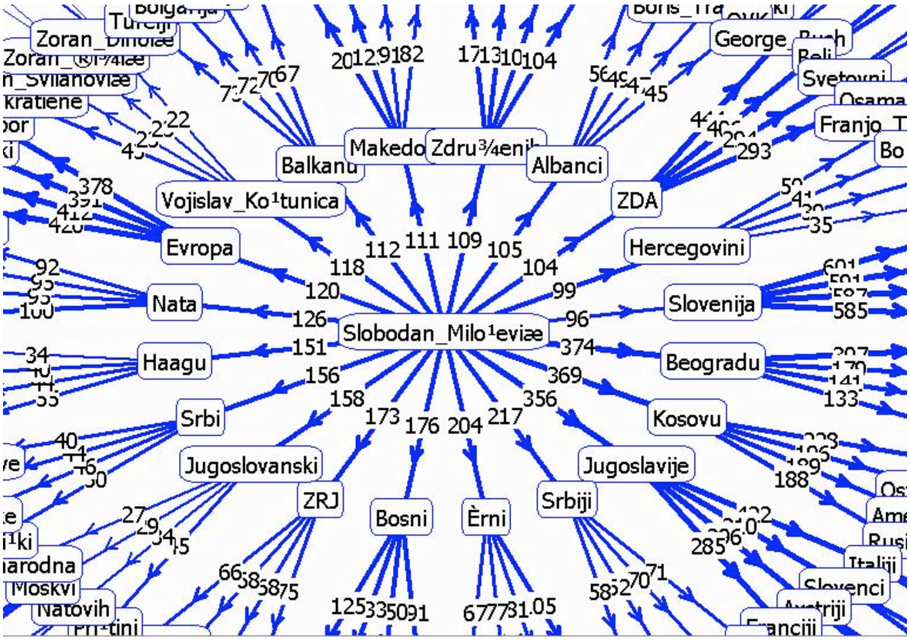

Fig. 1. Slobodan Milo_evi_ surrounded with the most frequent named entities from the 50.000 articles from the newspaper Dnevnik.

4 Results and Further Development

Both phases described in this paper, the methodology development and the crawler building, are being developed simultaneously. The methodological phase of the project is almost completed. As the theory sometimes differs from the praxis, we expect that some of the planned processes and implementation with the time should be corrected or changed.

The first version of the crawler we named WebBird was developed. At present we have tested it, thus capturing more than 500.000 web pages by domain from different types of organizations and analyzing their content and quality. Its real implementation is expected at the end of 2004. The analytic tools are still being developed.

NUK and IJS have received additional financial support for another project which is aimed at the development of a repository for electronic publications. In this regard we shall use the results of the methodology and specifications in the described project.

References

1. Abiteboul, S. … [et al.]. A first experience in archiving the French web. In: Maristella Agosti, Constantino Thanos (eds.) : Research and advanced technology for digital libraries: 6th European conference ; proceedings / ECDL 2002, Rome, Italy, September 16-18, 2002. - Berlin … [etc.] : Springer, 2002 (Lecture notes in computer science ; Vol. 2458). - pp. 1-15.
2. Brin, Sergey and Lawrence Page: The anatomy of a large-scale hypertextual {Web} search engine. *Computer Networks and ISDN Systems*, 1998, 30(1-7). - pp.107-117.
3. Brygfjeld, Svein Arne: Access to web archives: the Nordic Web Archive Access Project, paper presented at 68th IFLA Council and General Conference, August 18-24, 2002. URL: http://www.ifla.org/IV/ifla68/papers/090-163e.pdf (visited on 23 Mar 2004).
4. CEDARS: Digital preservation and metadata / Michael Day. *Sixth DELOS Workshop: Preservation of Digital Information, Hotel dos Templários, Tomar, Portugal, 17-19 June 1998*. URL: http://www.ukoln.ac.uk/metadata/presentations/delos6/cedars.html. (visited on 23 Mar 2004).
5. CENL & FEP (1999): *Code of practice for the voluntary deposit of electronic publications***.** URL: http://www.bl.uk/gabriel/fep (visited on 23 Mar 2004).
6. COMBINE: http://www.lub.lu.se/combine/ (visited on 23 Mar 2004).
7. Greenstone Digital Library: URL: http://www.greenstone.org/cgi-bin/library (visited on 23 Mar 2004).
8. Grobelnik, Marko and Mladeni_, Dunja: Efficient text categorization. In: [Proceedings of] *Text Mining workshop on the 10th European Conference on Machine Learning - ECML98,* April 21 - 24, 1998, Chemnitz, Germany. 1998.
9. *Internet Archive. URL*: http://webdev.archive.org/ (visited on 23 Mar 2004)
10. KULTURARW3: http://www.kb.se/kw3/ENG/Default.htm (visited on 23 Mar 2004)
11. Lynch, C., Preston, C., Daniel, R.: Using Existing Bibliographic Identifiers as Uniform Resource Names. - Los Alamos: The Internet Society,1998. URL: http://library.n0i.net/rfc/html/rfc2288.html (visited on 23 Mar 2004) .
12. Metadata for digital preservation: the CEDARS project outline specification. Draft for public consultation / The Cedars Project Team and UKOLN, March 2000. URL: http://www.leeds.ac.uk/cedars/MD-STR~5.pdf (visited on 23 Mar 2004).
13. MINERVA: http://www.loc.gov/minerva/ (visited on 23 Mar 2004).
14. Najork, Mark and Wiener, Janet L. : Breadth-First Search Crawling Yields High-Quality Pages (2001). In: *Proceedings of the 10th International World Wide Web Conference*,

Hong Kong : Elsevier Science, May 2001. - pp.114-118. URL: http://citeseer.nj.nec.com/najork01breadthfirst.html (visited on 23 Mar 2004).

15. NEDLIB: Metadata for the long term preservation of electronic publications / Catherine Lupovici, Julien Masanes, 2000. - (NEDLIB Report Series; 2) URL: http://www.kb.nl/coop/nedlib/results/preservationmetadata.pdf (visited on 23 Mar 2004).
16. Networked European Deposit Library (NEDLIB). URL: http://www.kb.nl/coop/nedlib/ (visited on 23 Mar 2004)
17. *Nordic Web Archive*. URL: http://nwa.nb.no/ (visited on 23 Mar 2004)
18. OAIS (ISO 14721:2002.): URL: http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html ((visited on 23 Mar 2004).
19. PADI PANDORA. URL: http://pandora.nla.gov.au/index.html (visited on 23 Mar 2004).
20. Page, Lawrence, Brin, Sergey, Motwani, Rajeev, and Winograd, Terry: The PageRank Citation Ranking: Bring. URL: http://dbpubs.stanford.edu/pub/1999-66/en, http://dbpubs.stanford.edu:8090/pub/1999-66 (visited on 23 Mar 2004).
21. Gatenby, Pam. Digital archiving - developing policy and best practice guidelines at the National Library of Australia. URL: http://www.icsti.org/icsti/2000workshop/gatenby.html (visited on 23 Mar 2004).
22. Preservation metadata for digital collections : exposure draft / National Library of Australia, 15 Oct. 1999. URL: http://www.nla.gov.au/preserve/pmeta.html (visited on 23 Mar 2004).
23. Preservation metadata for digital objects: a review of the state of the art. A White Paper by the OCLC/RLG Working Group on Preservation Metadata, Jan. 2001. URL: http://www.oclc.org/research/projects/pmwg/presmeta_wp.pdf (visited on 23 Mar 2004).
24. Žabicka, Petr: Archiving the Czech web: issues and challenges. In: Masanes, Julien, Rauber, Andreas, Cobena, Gregory (Eds.): *3rd Workshop on Web Archives, Trondheim, Norway, August 21st, 2003 : Proceedings*, pp. 111-117.