Harvesting and Archiving of Electronic Resources in Lithuania: towards Virtual Library

Remigijus Jodelis
LIBIS centre of Lithuanian National Library
Gedimino 51, Vilnius, Lithuania

Introduction

The rapid growth of the amount of electronic information in recent years has created the need for its long-term preservation. The digital environment in which this information circulates enables particular mobility and changeability of documents. This situation is even worse as the time between creation and preservation is decreasing because of constant technological advances. It has been calculated that the average life-span of a web document is 44 days. Besides, an electronic document can be easily corrupted without even the ability to recognize that the change has occurred.

The task of long-term preservation of electronic documents is most often taken by libraries which are striving to become more and more digital libraries. They have to preserve new types of objects such as databases, e-journals, e-books, websites as their original publishers are not always willing or capable of managing archival copies.

National libraries are in special position in regard to this consideration as they often have a special role assigned by state's legislation and also the necessary resources for creation of statewide archives of electronic documents. Many projects have been accomplished on international level also to facilitate cooperation in digital archiving and preservation.

In this paper I overview the concept of Web Archiving and the experience of the National Library of Lithuania in creating the Archive of Electronic Resources (AER). The purpose of this archive is to accumulate all electronic documents published in Lithuania plus those published outside which are related – for example, articles about Lithuania, resources created by emigrants etc.

Lithuanian Archive of Electronic Resources: project overview

In 2001 a project of the electronic Resources Subsystem in the Lithuanian Integrated Library Information System (LIBIS) was finished presenting a vision of the Archive of Electronic Resources, planning the stages of realization and necessary financial resources. The project is based on the model of the Deposit System for Electronic Publications and NEDLIB (the Network European Deposit Libraries) project documents. NEDLIB, funded by EU, is a collaborative effort of national libraries to construct the basic infrastructure for a networked European deposit library system for digital publications. It proposes a model for incorporating archives into integrated library systems and adoption of the Open Archival Information System Reference Model among other results.

The ER Subsystem is based on the same principles as other LIBIS modules and is aiming:

- to avoid duplicity of functions using the existing LIBIS modules and subsystems performing the analogous functions;

- to ensure one-time creation of bibliographic and authority records and their multifunctional usage, utilizing all information products of ER subsystem in other LIBIS structural subdivisions;
- to provide links between Archive of Electronic Resources and other LIBIS modules in the existing organizational-functional structure.

Creation of LIBIS Electronic Resources Subsystem as a virtual library will not only expand the capacities of the archive, but also products and services of LIBIS activities, including: current national bibliography, information supply, delivery of full-text electronic resources, and preparation of various bibliographic publications.

The legal base for the Archive of ER is provided by the decree of the Government issued in 1996 "Regarding the Order of Distribution of Legal Deposit Copies of Publications and other Documents to Libraries". Legal deposit includes books, brochures, periodicals, printed music, microforms, sound and video recordings, maps, fine arts, electronic resources and books in Braille. Since this decree doesn't define the coverage of electronic resources, selection principles for the inclusion into the Archive of ER have been created:

- dependence of ER to the Republic of Lithuania, which is defined by several indications – geographical location, which is identified by the domain "lt" in the Internet address, publisher's place of residence, author's domicile;
- status of ER – priority is given to the information products of official publishers and other organizations;
- all confirmed by publishers and officially registered editions of electronic resources.

The following resources are not to be accumulated:

- databases of internal usage of institutions
- unfinished and unofficial documents of separate individuals
- products of public communication on the net, e.g. e-mail, news discussion groups, informal communication, listings, internet games.

Types of Electronic Resources

There are different types of ER which should be treated separately and possibly by different subsystems of the Archive of ER.

1. The main part of electronic documents in the Archive would be those published on Internet in the form of more or less static web pages (also called "surface Web" documents). They can be accumulated using harvester techniques and will be the topic of the next section of this paper.
2. Dynamically created web pages linked to internal databases and other content of databases which is relevant to include into the Archive of ER (the so called "deep Web" documents). They are often behind firewalls and otherwise protected by publishers and thus not directly accessible. Accumulation of these resources requires agreements between a library and publishers based on Legal Deposit Laws.
3. Digital documents published on physical media (CD, MO disks etc.) They are much like traditional paper based publications with the exception that they require computer hardware and software for their utilization. The practical approach to archiving these documents in current situation is to store them in the original media while taking care of its life spans. A possible alternative may be to transfer the information to other storage forms (e.g. server hard disks) allowing easier access and preservation.

There are other forms of Electronic Resources including digitized manuscripts and other documents which however are out of the scope of this paper.

Web Archiving

The task of large-scale Web Archiving involves different solutions than those usually adopted by a library. The main difference is that web documents are out of any control. They usually are not properly marked with metadata, their presence on Internet may be terminated at any moment and the content may be changing constantly (the case of web portals). The situation is very little regulated by legislation and only the official publishers may be subject to the Legal Deposit Law.

There can be various levels of state-wide Web Archiving – one may want to collect all available documents in the national domain some related outside, or to perform a selective acquisition with various goals and selection criteria. The decision about the form of Web Archiving is heavily dependent on available technical resources.

The most common approach to Web Archiving is to use harvesting. Web harvester is a system which downloads web pages, extracts new URLs from HTML code and includes them into the list for further downloading if they are not already done. This way robot software searches the selected web space and makes a copy of every available document on the server hard disk. A similar approach is used by popular web search engines (Google, Lycos, Yahoo! etc). The described process allows after specifying several successful starting URLs to collect a large part of specified web space. It however dos not guarantee that every available online resource will be gathered. Therefore the harvesting process requires monitoring and possible addition of new URLs not reached by the robot.

There are several software packages that are used by various countries for general or selective harvesting of their web space. The most used in Europe are NEDLIB and COMBINE harvesters. They are capable of large-scale harvesting and are robust enough to be exploited continuously. NEDLIB harvester has been further developed by Helsinki University Library and is freeware. One of the biggest advantages of this harvester is that it has been adapted specifically for Web Archive creation.

Another approach to archiving of web documents is to cooperate with producers and publishers so that they submit their publications to a library. It may be done periodically as the eb site is updated. The base for such cooperation can be the Legal Deprosit Law or it‘s equivalent, so that the whole process is similar to the acquisition of paper documents. The obvious drawback of this approach is that only officially registered publishers are involved and a lot of other publications would not enter the archive. Among the advantages is the opportunity to obtain high quality metadata created by publisher which would facilitate cataloging of archived documents.

Harvester problems

There are a number of problems when one tries to use only automatic web harvesting method. Existing web harvesters are not designed for flexibility and it is difficult to arrange scheduling of harvesting times because they do not distinguish between different categories of web sites.

Another aspect of harvesting is how to assure that web documents are collected in their entirety, with all linked parts intact and working when they are accessed from the archive.

This is particularly difficult with dynamic web pages containing JavaScript mini programs and other embedded software. Our experience with NEDLIB harvester shows that large number of dynamic web pages are not handled properly and this occasionally even results in gathering only a small part if not just the starting page of a selected web site. Ideally a harvester should have a Java VM to be able to parse HTML page content properly but this is hardly achievable in the nearest future.

Closely linked to previous is a problem of dynamically created web pages and URLs with parameters. A harvester is actually interacting with a web server when attempting to collect this type of documents. In pathological cases it may be getting new URLs forever from the same server. It may be therefore dangerous to configure a harvester to gather parameterized URLs despite the fact that a considerable part of the web space is occupied by dynamically generated content.

The conclusion to draw from these considerations is that today's harvesting software is still too inflexible and may be suitable only for acquisition of static online documents – essentially stored files.

Web harvesting results in Lithuania

It was decided in the project of the ER subsystem of LIBIS to utilize the NEDLIB harvester for the purpose of Web Archiving. The scope of harvesting is to collect all available data in "*.lt*" domain and related web sites. A list of Lithuanian web sites in "*.com*", "*.net*" and other domains was collected consisting of about 300 items. We decided to use the popular web directory "Lithuania Online" along with few other big websites for starting URLs of the harvester.

A server in the disposal of the National Library of Lithuania was appointed with the following characteristics:

HP NetServer E40	
CPU	2x400 MHz
RAM	768 Mb
HDD	3x80 Gb
OS	Linux RedHat 7.3

NEDLIB harvester was installed and successfully tested. The harvester keeps its internal data of documents and URLs lists in a MySQL 3.23 database. It was configured to leave out URLs with parameters so that only static documents should be attempted to collect. Harvested documents are packed with *gzip* after finishing harvester daily session.

After good initial results thee first harvesting cycle was started in 25 October 2002. In three weeks of daily operation a large portion of accessible documents was estimated to be covered (see statistics at the end of this document).

Another copy of NEDLIB harvester was started in 27 November 2002 which was designated to collect periodic web publications on a selective basis. About 60 URLs of officially registered e-journals or electronic copies of periodicals were listed together with their publishing periodicity. Additional software was developed to ensure their timely harvesting in regard to the fact that NEDLIB harvester is not capable of variable scheduling.

As mentioned above, a large number of websites (especially e-journals) are dynamically created web pages. Considering this it was decided for fuller representation of Lithuanian web space to make efforts in obtaining priority electronic publications directly from publishers. The National Library of Lithuania prepared a contract on acquisition of e-publications and copyrights which was sent to main publishers in March 2003. Today we already have first contracts signed and first documents together with metadata starting to arrive.

Generally results of web harvesting are quite satisfactory. First of all the NEDLIB harvester has showed very good operational characteristics. It hasn't crashed since the start and is able to manage very large quantity of data including it's ~1GB size MySQL database. The size of archive in December 2002 was as follows:

Total urls retrieved:	2 089 943	
Reharvested:	612 600	
New urls:	1 477 343	
Out of *.lt* domain	4 982	(0.3%)
Parametrized urls:	96 501	(6.5%)
Total size of documents:		**75.5 GB**
Average size of metadata per URL:		142 bytes

Since the start in October 2002 it has made four harvesting cycles. Because the amount of database increases considerably with each full cycle of harvesting, increasing archive access times, it was decided to perform further general harvesting 2 times a year. Of course, gathering of periodicals will be performed constantly as it does not require large storage costs.

Another direction in which the development of the Archive of ER in Lithuania takes place is indexing and cataloging of collected information. Only the resources meeting selection criteria are taken and extensive bibliographic records are prepared by catalogers in UNIMARC format. These records are included into the National Bibliography Database and are available via OPAC. For the rest of documents not reflected in the NBDB only short bibliographic records are being created by automated conversion software from the metadata found in the documents. As very few HTML pages in the Web have quality metadata, these records are not very valuable for the users. To improve the situation it was decided to recommend publishers to use *Dublin Core* metadata standard for marking online and other electronic documents. Also a template was created to facilitate the generation of metadata in this format.

Considering the full-text indexing of harvested documents the solution has yet to be found. Some indications show it worthwhile to pay attention to web indexers used in Internet search engines (like *mnogoSearch* etc). Of course, full-text indexing will be implemented in the future.

Development steps of the Archive of ER

While the Archive of Electronic Resources of Lithuania has been created and is working, it is worthwhile to foresee further developments towards fully operational digital library. The next logical step after creation of the Archive is to provide access for library users. First of all a visitor must be able to search bibliographic records. For most of the documents in the AER

only short records will be available consisting of URL of harvested document plus metadata found in the document. The persistent unique identification number must be assigned to each document allowing consistent reference.

Next, the access to AER may be organized through the same interface as internet – namely browser interface. This was already made in the Lithuanian National Library but only for catalogers yet. Dynamic web pages are generated by web server operating on the same machine as both harvesters. Archive contents are retrieved by their identification number or by URL and harvesting time. The ongoing work is being done to improve the interface and to make it more comfortable for users of the Archive.

Also, considering copyrights and adjacent authors' rights it has been decided not to allow copying or printing of archived document. These restrictions may be relaxed in future for particular website copies as agreements will be made with their publishers.

Next stage of the project will be accomplished after creation of full-text index of the archive content. It will greatly improve its value enabling search by keywords. Finally the Archive of Electronic Resources will be integrated into the website of the National Library. The website already has access to databases of bibliographic records and additional interfaces with full-text document retrieval capability would create a truly digital library. Still, a lot of work remains to be done including solutions about copyrights protection, efficient retrieval of information from the Archive and inclusion of other types of electronic resources.

References

Hodge, G. (2002). Archiving and Preservation in Electronic Libraries. RTO IMC Lecture series on "Electronic Information Management for PfP Nations".

Hallgrimson Th. (2003) Survey of Web Archiving in Europe. (to be published)

Varnienė R. (2001) Selection, accumulation, usage and archival storage in the Lithuanian Archive of Electronic Resources: methodics of organizational and technical means. Vilnius, 2001. (in Lithuanian)

Bulavas,V. and Varnienė R. (2002) Lithuanian Publishing and Bibliographic Control. Vilnius, 2002.

Arms, W.Y., Adkins, R., Ammen, C., Hayes, A. Collecting and preserving the Web: the Minerva prototype. [http://www.rlg.org/preserv/diginews/diginews5-2.html]

The PANDORA project: a summary of progress. [http://pandora.nla.gov.au/]

Supplement: Statistics

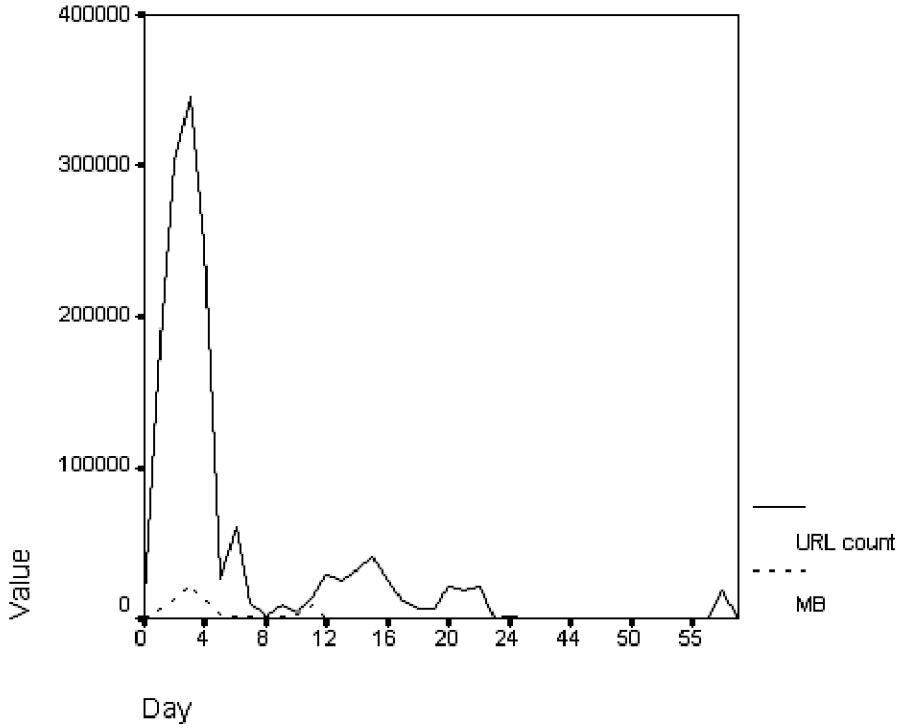

Chart 1. Harvester activity since start: downloaded new URLs count and content size (in Megabytes).

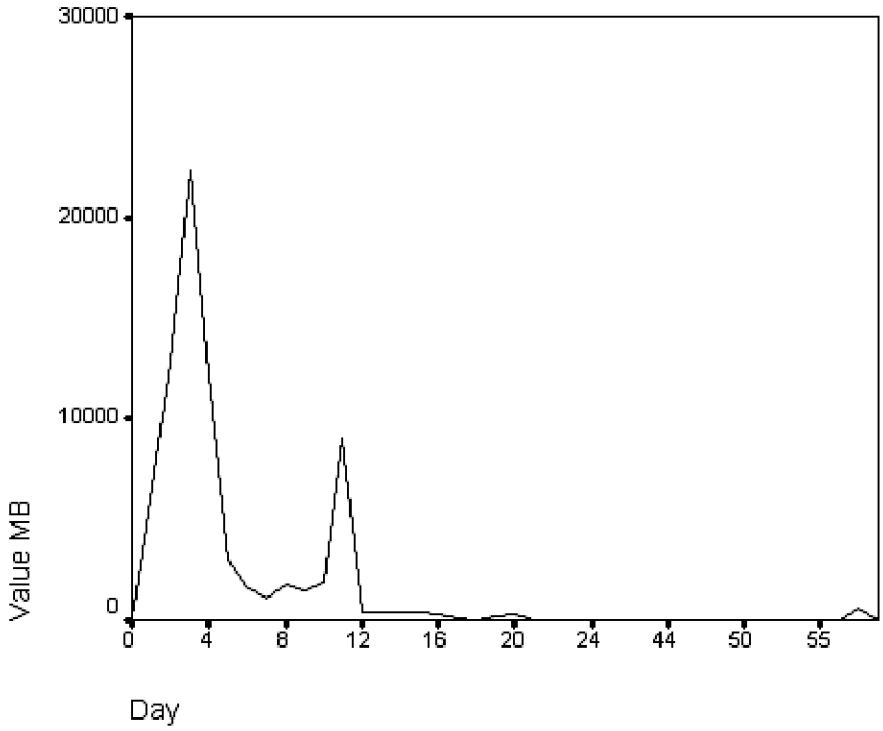

Chart 2. Harvester activity: downloaded content size (in Megabytes).

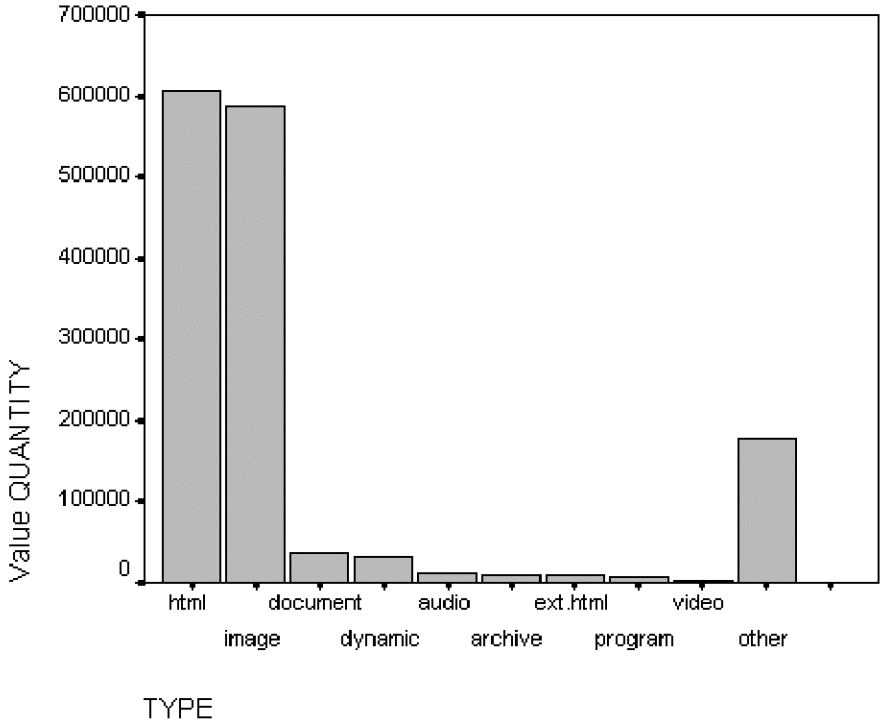

Chart 3. Distribution of archived documents by their types.

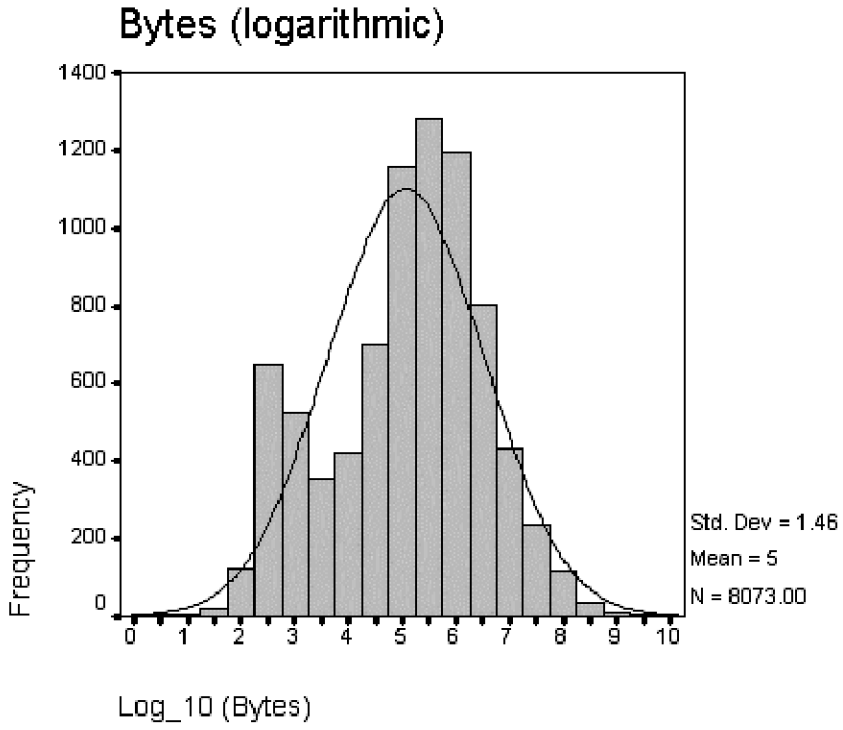

Chart 4. Histogram of host size in bytes (in logarithmic scale).