Part of Our Culture is Born Digital - On Efforts to Preserve it for Future Generations

Andreas Rauber, Andreas Aschenbrenner
Department of Software Technology and Interactive Systems
Vienna University of Technology
Favoritenstr. 9-11 / 188
A-1040 Vienna, Austria
www.ifs.tuwien.ac.at

Abstract

The Internet has already established itself as one of the key information and communication media of our time. With Internet-specific types of information representation and art evolving, and with social groups forming in cyberspace, the Internet and its content is turning into an important part of our cultural heritage. This new digital cultural heritage is deemed worth preserving by an increasing number of initiatives, aiming at providing access to our contemporary Internet for future generations.

Numerous challenges have to be addressed in the creation of any such archive, concerning the acquisition of data, its long term storage and preservation, as well as access provision. In this paper we provide an overview of the most prominent challenges encountered with respect to archiving the Internet and discuss possible solutions. We furthermore present a report on AOLA, the Austrian initiative of creating an on-line archive.

Keywords: Cultural Heritage, Preservation, Emulation, Migration, Internet Archive, AOLA

1 Introduction

With the rise of the Internet, digital information has become seemingly ubiquitous as a medium for communication and expression [LK98]. With the amount of information being made available in digital form, the need to archive and preserve access to these documents becomes an essential asset, which is only starting to be appreciated to its full extent. Organizations like the *Long Now Foundation* [1] with its *clock/library* projects are seeking to promote slower and better thinking and to focus our collective creativity on the next 10.000 years, pointing out the need to preserve the material being created [Ban00].

Nevertheless, one might argue that a high percentage of the data on the Internet is junk, useless, or even misleading information hardly anybody actually looks at. Furthermore, most of the information that is worth to be kept for future generations, is available

[1] http://www.longnow.org

in traditional media, anyway. Texts, music, films, and pictures all have their traditional physical media.

However, with the new possibilities offered by digital media, new forms of information representation emerge. The expressive power of hypertext documents with their non-linear link structure, as well as multimedia documents integrating video, sound, and interactive components, cannot be adequately represented in traditional forms. Furthermore, apart from the mere information contained in the documents, the network formed by cross-referenced documents offers another dimension of information largely unmatched in conventional media.

With the popularity of the Internet rising, the emergence of new communities adds a social component to the web. When the Internet used to be populated solely by a few scientists and geeks, interaction mechanisms and the available content were limited and adjusted to their needs and habits. Yet, the type of information made available as well as the way it is presented changes drastically as new groups of users discover the Internet, using it and adapting it to their needs, forming and evolving into new and diverse communities.

This is only the beginning of a process where our cultural heritage increasingly incorporates digital forms. It results in a shift of focus, making the archiving of digital material not primarily a question of storing asserted facts and figures, but of preserving the big picture, of capturing the sociological and cultural dimension of the Internet and its "inhabitants". Already now the material gathered by spear-heading initiatives has proven a useful resource to historians analyzing e.g. elections [Kah97]. As we start to recognize the importance of preserving at least parts of the entities that make up web, more and more projects are initiated to address the technical challenges associated with such an endeavor.

In this paper we provide an overview of issues pertaining to the archiving of digital material, with a special focus on the Internet, specifically the World Wide Web. To put it in a nutshell, these can be summarized as (1) acquisition, (2) storage, (3) preservation, and (4) access provision. First of all, the necessity to capture the characteristics of the web, i.e. to obtain the documents, their content, look-and-feel, as well as their role within the larger network of interlinked information poses serious challenges. The vast amount of material available on the web, the lack of a central listing, as well as the volatility and frequency of change call for carefully designed methods of data acquisition to meet the goals of building an Internet archive.

Secondly, once documents have been collected, long-term storage is a tacit requirement for any archive, yet demands careful consideration. With the current pace of technological changes, storage media become obsolete every few years, and so does the technology required to read the archived material. This requires migration strategies to move data from one storage system to the subsequent generation. A similar migration step is also required when the physical life time of a storage medium is reached, independent of the life span of the technology as such.

However, the most serious challenge and threat to the archiving of digital material lies within the concept of digital objects themselves. Contrary to conventional books, or even ancient stone plates, digital objects cannot be read as they are. Rather, they require special software to access and interpret them. Yet, the rapid pace at which these technological changes proceed may render most documents unreadable within a few years, as the original software is replaced by new, incompatible versions. Unfortunately, even archiving the software necessary for interpreting digital artifacts does not provide a

solution. Software typically is designed for specific hardware platforms, i.e. a specific type of computer system, consisting of a specific set of chips. These systems would need to be preserved as well, resulting in a museum of outdated hardware nobody is able to maintain. Several approaches are pursued to tackle this problem, such as emulating obsolete systems or converting the digital objects into formats that can be interpreted easily even on future systems.

Last, but not least, once such an archive has been created, providing access to its content poses additional challenges, which to a large degree depend on the needs of its users and their goals, be it locating a specific document, or analyzing the evolution of social groups in cyberspace. This again has a strong influence on all aspects addressed so far, i.e. the information acquired for archiving, the storage media containing the archive and its means of access, as well as the form in which the documents are made accessible.

Obviously, legal issues merit special consideration, with any such project touching on privacy, copyright, as well as ethical questions. While hardly any congruent legal framework for the creation of Internet archives exists, there is a general agreement that web sites may be deemed as published - and thus publicly available - material that may be archived. Yet, providing access to the documents stored in the archive constitutes a new situation. Thus, no project currently provides open access to its archives apart from special research projects. Due to the complexity of these legal issues, we are not able to cover these aspects in detail within the scope of this paper. We rather refer to [Lan96] for detailed coverage of this topic.

The remainder of this paper is structured as follows. We start with a presentation of the major technical challenges related to the creation of Internet archives in Section 2. This is followed by a review of related projects in this field in Section 3. In Section 4 we present the Austrian On-Line Archive AOLA, the Austrian pilot study undertaken as a joint cooperation between the Department of Software Technology and Interactive Systems and the Austrian National Library. After a review of the results of the pilot phase, some conclusions are presented in Section 5.

2 Challenges

Numerous obstacles have to be tackled when building an archive of the Internet. They range from the acquisition of the material, via storage and long-term preservation of the data, to the provision of access. In the following sections we will take a more detailed look at these challenges, pinpointing the problems and presenting approaches that are being pursued.

2.1 Acquisition

In terms of acquiring the documents to be included in the archive, three important decisions have to be made. First of all, the scope of the archive has to be determined, such as limiting the archives' content to a nation's web space, with the question of what actually constitutes the national web space being trickier than might be imagined at first glance. Secondly, the actual documents to be included and how they are acquired has to be determined, with possibilities ranging from careful manual selection to automatic harvesting, each with serious consequences and implications. Last, but not least, given the volatility of Internet documents, the frequency at which documents are entered into

the archive gains significance.

Assuming an archive embarking on a national scope, it has to be constituted, where the borders of the national web space are drawn. Obviously, all sites being part of the very country's national domain (*.at* in the case of Austria) are within that context. Yet, many servers located in a country are registered under a foreign domain, most notably under domains such as *.com, .edu, .org*, but also under "foreign" national domains, such as *.cc*, or *.tv*. While the logic addresses of these domains have no association with, e.g., Austria, the sites themselves might still be physically located in Austria and be operated by Austrian organizations (e.g. *www.austria.com*). They thus most probably are considered worth to be included in a national archive of Austria. Last, but not least, web sites dealing with topics of interest, such as foreign web sites of, e.g., expatriate communities, or other sites dedicated to reports on Austria (so-called *"Austriaca"*), should possibly be collected, even if they are physically located in another country.

The next decision to take concerns the documents to be included in the archive and the way how they are obtained. Basically, there are two different approaches, which can be described as (1) manual selection in contrast to (2) automatic harvesting [MAP00].

The first approach requires all material to be manually selected or relying on donations from various organizations. All digital documents would then be scrutinized in order to decide, whether or not they are worth to be stored. Thereby, a consistent, carefully sorted archive could be built. Yet, a comprehensive solution can hardly be attained.

On the other hand, simply anything one can get a hold of could be stored in the archive. Gathering material this way can be achieved in an automatic manner. Starting from a number of sites, so-called web crawlers, such as those used by current Internet search engines, move to other sites following the links they find. Due to the highly interlinked structure of documents on the Internet, these crawlers are able to harvest a considerable portion of the web. Yet, sites, which are not part of the initial setting and are not linked to from any site, will not be collected. Thus, any solution will never be complete but only far-reaching.

The latter approach is to be favored for several reasons. Obviously, any more sophisticated selection from millions of web pages would call for heavy personnel inputs. A comprehensive solution could probably never be reached. More important, however, it is hard to distinguish between what in the future will be considered valuable and what worthless. Historians working with newspapers preserved from a hundred years ago assess obituaries, advertisements, and other sections commonly considered junk very interesting. If there had been a selection upon this material, we would never rejoice in possession of this valuable source of information. Of course, the Internet comprises loads of "Sex and Crime", but - whether we like it or not - this is part of our present culture. Nobody could sincerely claim being able to judge upon what will be considered valuable in the future.

Handling the gathering of the material to be stored in the archive in an automatic fashion also raises a legislative problem. On the Internet there are sites publicly available, that are prohibited by law, such as Nazi propaganda or child pornography. Since automatic crawlers are not able to discern these illegal sites, they will include them in the archive along with all the other documents. Thereby the archive may contain offensive material without being aware of it. This topic has to be handled with care and sensitivity in terms of access provision.

Data in the Internet characteristically has a high volatility. Estimates put the average lifetime of a document at 44 days [Kah97]. Therefore, it is inevitable that intermediate versions of some documents are missed out, most actually will be lost at all. This again

impedes the striving for a solution as complete as possible. Reconsidering the initial motivation of wanting to convey a picture of the Internet at a given time, failing to archive each and every document turns out to be not that severe a loss. After all, the main goal is to give an impression on the look-and-feel of the Internet in our days, which can be achieved even if a certain percentage of material is lost. Taking this one step further, it calls into question the continuous downloading of each new or changed file that can be found. Actually it should suffice to make a snapshot of the Internet at certain intervals, say, every half a year. It might be desirable to adjust this frequency selectively, as the volatility of certain web sites differs from others. For instance, newspapers change more often than the average web site. Perhaps it is desirable to have an extra snapshot of political sites during election times. In general, however, collecting material periodically facilitates the task without having a negative effect on the result.

Most initiatives currently in action embark on a largely automatic strategy. Recurring snapshots are considered at half-year intervals. For this task crawlers are used to collect virtually anything they come across, provided it belongs to the predefined scope. Resuming the assumption of trying to acquire a country's national web space, we have three lines of interest as pointed out previously. Documents registered under a nation's domain can easily be identified. In contrast to this are hosts registered under a foreign domain, but located in the very country, which may be identified with some additional effort using Internet registries. Sites located in another country, but dealing with topics of concern cannot be recognized automatically. An understanding of the contents is required to identify those, which cannot be done automatically with state-of-the-art technology. A limited automatic classification based on heuristics is a task worked on, but at this point of time all these sites have to be defined manually.

Various ways of interaction with the user and dynamically generated web sites are on the increase. To give an impression on how the trail of navigation continues after an interaction form, i.e. what an individually generated response to an interaction would look like, a means of handling these dynamic web sites must be developed. This is not possible in an automatic manner up to now. Again, semi-automatic ways of retrieving a few representative probes to convey the impression are being worked upon.

2.2 Archiving for long term storage

All the collected documents obviously need to be stored. Even though tremendous amounts of data are involved, the sheer amount is not the main challenge to be faced. Admittedly, the Internet is growing at a very high rate. At the same time, however, storage media is getting more powerful. A far more serious problem when archiving electronic documents is to guarantee long term storage.

The CD-ROM, one of the most popular backup storage media nowadays, has a predicted physical lifetime of 50 to 100 years provided normal environmental conditions before it degrades and looses the data stored on it [Bog96]. Yet, considering the young age of the CD, there is, obviously, no practical experience on this subject. A somewhat limited lifetime applies for any technology. Therefore, the functionality of the archive has to be maintained by periodically copying the data to new storage media before the old deteriorates beyond the point at which the information can be retrieved.

This tackles another problem at the same time: Since technology standards are evolving at an enormous rate, the type of storage media has to be kept up-to-date. While CD-ROM's might survive and store data for a hundred years, chances are that no drives

capable of reading CD-ROM's will be available by that time anymore, as new forms of storage media will by long have replaced it. New media for storing digital information rapidly replace older media, and reading devices for these older media become no longer available. We just have to imagine how many types of storage media we have seen passing in the last 30 years, starting from punch cards, via old magnetic tapes, various types of hard disks, to 8 inch and 5 1/4 inch floppy disks. A national library cannot become an IT museum with hardware and software, which, in the end, no one is capable of operating. Indeed, technological obsolescence represents a far greater threat to information in digital form than the inherent physical fragility of many digital media [MA96].

Currently, tapes, which offer a lot of storage space at low cost, are used most frequently for long-term storage. However, tapes typically have rather large access times requiring considerable time to load a tape from a tape robot and winding it to the correct position before a file can be retrieved. Hard-disk arrays provide fast access to considerably large amounts of data. Although they are also rather expensive, the price difference to tapes is decreasing for comparable amounts of data. This causes many initiatives to move from tape-based archiving to hard-disk arrays to allow convenient access to the data, keeping tapes as back-up storage media.

Further aspects of data storage requiring serious consideration concern the protection of the archived material as well as detailed cost vs. risk analysis and trade-offs. Projects thus frequently will opt for distributed replication of their data, spreading it across different sites in different regions to escape natural disaster. Furthermore, quality analysis as well as precautions against data loss have to be taken to guarantee a high probability of correct data preservation. For more details on these issues, refer to [CGM01a, CGM01b].

To avoid technological obsolescence the data is migrated periodically to new storage media. This has not been done repeatedly up to now, since such archives have emerged only recently.

2.3 Accessibility and interpretation of digital objects

Copying the data to new storage media offers a solution for long-term storage of digital material. Yet, changing technology standards are not only a hardware problem, this applies to software at an even larger scale [Man00]. Acquiring access to information stored in electronic documents requires software that interprets the corresponding digital object. This characteristic of new media objects constitutes one of the most important differences to conventional media. While ancient stone plates, papyrus scrolls, and codices can be accessed as long as their physical being is preserved, additional tools are necessary for gaining access to electronic media. This applies not only to our modern Internet documents, but also to more conventional media such as audio or video data. Software programs in turn need a specific type of hardware platform they can run on. They not only follow the hardware development, though, but evolve themselves. New software applications continue to emerge, superseding others. One and the same program changes the data format to extend its functionality. For these reasons data formats exist at a vast variety, and even standard formats are replaced and supplemented by new. We, hence, ever so often end up with not so old digital documents that, although being stored in perfect condition, cannot be accessed and read by any program available on our state-of-the-art computer system. Even if we have the original software at hand, chances are that it will not run on our current system, as it was designed for older systems that are no longer available.

Considering the still relatively young age of computer science it is not surprising that long-term preservation of data is still subject to massive research. There is nothing such as a unique and sound solution to this task, and there probably never will be such an ultimate solution. Various strategies for dealing with technological obsolescence are being followed, and reference models, such as the Open Archival Information System (OAIS) [CCS01] are being developed. Apart from the infeasible endeavor of trying to maintain a museum of obsolete computer systems, two realistic approaches can be identified. With the *migration and conversion approach* data objects in the archive can be kept up-to-date by periodically converting them to standard data formats that are in use at that time. Secondly, it is possible to *emulate* obsolete hardware environments on new systems. Subsequently, data formats can be interpreted in the future by software in use now. As the right approach, if there is something like that, can only be identified as such in the future, a combined solution is probably the most fruitful. We will take a closer look at the characteristics of both approaches in the following two subsections.

2.3.1 Migration and data conversion

The migration of digital information refers to the periodic transfer of digital material from one generation of computer technology to a subsequent. Please note, that we can basically distinguish between two different types of data migration. The first type, detailed in Section 2.2, refers to the migration, i.e. copying, of data from one storage media to another, without any changes to the document itself. It primarily counters the technical volatility of storage media, as well as their physical decay. The second type of data migration is better described as *conversion*, as it refers to the transformation from one data format to another, actually modifying the digital object. Both types unfortunately are simply referred to as "migration" in the literature.

As the variety of different data formats is vast and will be so in the foreseeable future, documents are converted into few, selected standard file formats. A number of standard data formats, each preserving a particular characteristic of documents, needs to be identified. These, for instance, may be pure ASCII text or XML to provide content-based access, and Postscript or the Portable Document Format (PDF) to convey the looks of a document. Other formats are required for image, audio, and video data, while a series of snapshots might preserve some characteristics of dynamic documents. Consequently, all data stored will be in few basic formats that can be interpreted by a small set of standard software at any time.

Taking into account that only a few frequently used standard formats are used for storage in the archive, it can be expected that converters to a newly developed, superior standard type will be available. This is due to the tremendous amount of files existing in a standard file format, all of which require to be converted from the then obsolete to the new standard format. Needing only few converters and at the same time the prospect of having those at disposal easily facilitates a stable and manageable maintenance process. Even more important, fast and low cost access to any document can be provided, as the standard format, in which a document is stored, is readily readable by any computer system. No special transformation or the installation of any particular type of software is required.

The complexity of the migration process will depend on the nature of the digital resource, which may vary from simple text to an interactive multimedia object. However, converting data to another software format entails a loss of functionality. Cross-references,

indices, interaction mechanisms, automatic updates, formulas and the like will usually be lost during the conversion stage, since the particular standard file format will not support the same functionality as the original software. Furthermore, the authenticity of the original object is thereby corrupted. In other cases, as for example for some forms of interactive art on the Internet, converting to another format without loosing all of its characteristics is not possible at all.

For an immediate solution, migration appears to be very flexible and cheap. In the long run, however, it can be time-consuming and costly. Another drawback poses the difficulty to predict the frequency at which digital information will need to be migrated, as well as the fact that the amount of data to be converted continues to grow during the life-time of the archive.

2.3.2 Emulation

To tackle the disadvantages of the migration strategy, an emulation approach may be followed. Emulation refers to the process of mimicking, in software, a piece of hardware or software so that other processes think their familiar environment is still available in its original form [Fee99]. Consequently, digital documents can be kept without being altered, thus maintaining the integrity and its original look-and-feel. Together with the documents the access software is archived. It will be run on a future system which simulates a contemporary computer environment without the software noticing, giving access to the stored digital documents.

Emulators for hardware platforms are created as the necessity rises. This, of course, gives the impression that the actual work is only shifted to a later point of time. Exacerbating this is the fact that the development of specialized emulators is very expensive and time-consuming. Still, theoretically emulation is the most stable model and a conceptually clean solution. In fact, if preserving the original functionality, recreating the look-and-feel of a document is a prime objective, it is the only reliable way [Rot99].

Vital for this strategy is the keeping of meta-data which identifies the exact environment needed for running the stored software and thereby providing access to the documents. Therefore, a precise technical specification of every computer system to be emulated has to be obtained. As the emulators will be written on an unforeseeable computer system of the future, it appears hard to predict what information will be needed.

Emulation approaches are analyzed in detail in the *CAMiLEON* project [Gra00]. A different approach using a Universal Virtual Computer (UVC) and distinguishing between emulators for programs to be run as well as data files to be accessed is described in [Lor01]. There are many different attempts to overcome these the described problems. While its applicability has so far only been demonstrated in a few situations, emulation nevertheless is a promising approach.

2.4 Using the Archive

The interest in digital cultural heritage has emerged only recently. Thus, many countries still lack a legal framework. Besides a revision of the deposit law, copyright and privacy legislation have to be taken into consideration [Man00]. As the archive consists of publicly available documents only, it is in principle not violating the copyright law. Yet, there are companies that would object to the creation of such an archive. For example newspapers that have their daily edition freely available, but charge money on articles from their own

archive, could fear a serious loss of income, as people would have the possibility to obtain the documents from other sources.

However, these issues are only of minor concern. The reason for this lies in the characteristic of any such archive and its inherent incompleteness. As we have seen in Section 2.1 it is impossible to archive *all* data on the Internet. Most data will not be included in the archive anyway, given the short average life span of a document, as snapshots are taken at intervals of half a year or more. Furthermore, special means of access can be negotiated, including limited access to specific domains, or providing public access only after a certain period of time. If privacy concerns are at stake, a right to have specific documents removed from the archive may also be granted. Most robots used for data archiving further obey the so-called robot-exclusion files. These files allow the owner of a site to specifically exclude certain parts or all documents of a site from being harvested. As a consequence, only documents that are allowed to be harvested by an automatic crawler will be included in the archive. Yet, most of these issues still have to be covered by national deposit and copyright laws.

Another aspect of the copyright law must not be forgotten. The authenticity of the data has to be guaranteed. By developing means to electronically sign the data just after it has been collected, the authenticity could be examined at any time. Coupled with an open-source approach, where anyone can follow the way from collection to signing and archiving, this should settle any doubts. However, the question of authenticity with respect to format conversion still has to be addressed.

As for granting access, an interface to the archive could look similar to regular navigation tools for the Internet. Having added another dimension, the user has the possibility to go back and forth in time. Thereby the evolution of a web site can be followed. Like in the regular Internet an indexing and search function could be provided.

Since the archive has to deal with enormous amounts of data, a lot of storage space has to be supplied at a preferably low prize. This is only achieved by slow media, which can be concealed to a certain extent by hierarchical storage and a powerful hashing scheme. Furthermore, in an emulation approach, a specific emulator may have to be created before the software required to access a specific document can be installed and run. For these reasons, undoubtedly work is impeded because of slow access.

3 Related Work

Due to the importance and urgency of issues pertaining to the archiving of Internet content, numerous projects have been started, that address various aspects of the numerous challenges in this field. In the following subsections we provide an overview of the most prominent amongst these, outlining their goals and general principles, as well as reporting on their current state.

3.1 The Internet Archive

The biggest initiative is the *Internet Archive*[2], a public non-profit commercial venture. Located in the Presidio of San Francisco it was founded in 1996 to build an 'Internet library', with the purpose of offering free access to historical digital collections for re-

[2] `http://www.archive.org`

searchers, historians, and scholars. In March 2001 the archive's collections comprise more than 43 terabytes.

Collection is not carried out by the *Internet Archive* itself, but it receives the material from third parties. The main contributor is *Alexa Internet*[3], a web navigation service.

Alexa's robot gathers more than 100 gigabytes of data a day from all around the world. These huge amounts of data are not transferred to the archive until the material in them is at least six months old. All publicly accessible World Wide web pages, the Gopher hierarchy, the Netnews bulletin board system, and downloadable software are included in the material acquired. The content of the collection is by no means filtered or selected. For this reason the *Internet Archive* states explicitly, that the collections may contain information that might be deemed offensive, disturbing, pornographic, racist, or otherwise objectionable, not to mention the accuracy and completeness of the information.

No copyright laws are violated, as the collection consist of publicly available documents only. If there is any indication that their owners do not want them archived - such as robot exclusion mechanisms - they are obeyed. Even manual removal is possible on request.

To guarantee long term preservation of the collection three levels of actions are undertaken. Maintaining copies of the archive's collections at multiple sites alleviates the risk of accidents and natural disasters destroying the data. Over time storage media can degrade to a point where the data becomes permanently irretrievable. Therefore, the data is migrated to new storage tapes frequently. These two means counteract the vulnerability of the hardware. As advances are made in software applications, many data formats become obsolete. Software and emulators are collected that will aid future researchers, historians, and scholars in their research.

The *Internet Archive* makes the collections available at no cost to researchers, historians, and scholars. At present a certain level of technical knowledge and programming skills are required to be able to access them. Some of the historical and scholarly users of this data have been the Smithsonian Institution, Xerox PARC, AT&T Labs, Cornell, Bellcore, and Rutgers University. Uses have been the display of historic web sites, the study of human languages, the growth of the web, and the development of human information habits among others.

3.2 NEDLIB - Networked European Deposit Library

One of the most comprehensive approaches has the *Nedlib* project[4] covering in detail aspects of Internet archiving. *Nedlib* is a collaborative project consortium, headed by the National Library of the Netherlands and including eight other European national libraries, a national archive, and three main publishers. Funded by the European Commission's Telematics Application Programme it was launched on January 1st 1998. Officially the project ended January 31st 2001, yet, work on it is still continuing.

The main goal is to find ways to preserve access to both on-line and off-line (physical format) digital publications. *Nedlib* aims at constructing the basic infrastructure upon which a networked European deposit library can be built. Its objectives concur with the mission of national deposit libraries to ensure that publications of the present can be used now and in the future, extending their deposit to digital works. Those efforts are directed towards converging solutions and thereby contribute to an emerging infrastructure for digital deposit libraries [WD99].

[3] `http://www.alexa.com`

[4] `http://www.kb.nl/coop/nedlib`

Nedlib provides a forum for the exchange of best practices and serves the purposes of consensus building and spreading research costs. Not a stand-alone monolithic system is proposed, but a "plug-in" model in which the deposit system is embedded in existing or emerging digital library infrastructures. Therefore, *Nedlib* has no single philosophy in collecting, storing, providing access, and preserving the digital material. It offers a model, a framework flexible enough to accommodate the needs of all the participants, whilst they follow their own approaches. For example, in preserving access to the digital objects for the future, format specification to be able to interpret the stored data, meta-data in general, migration to new formats, and hardware emulation are dealt with, thus taking an universal look at the issue. Guidelines and technical standards make exchange of knowledge possible.

Some tools were developed within the project, such as a special *Nedlib Harvester*, an application for harvesting and archiving web resources. It is maintained jointly by Helsinki University Library and the Finish Center for Scientific Computing. The German National Library has developed a system for multimedia access, an integrated client/server environment to support the workflow for electronic publications. And there are likely to be more tools coming up as the participants build up their digital deposit libraries.

3.3 Kulturarw3 - The Swedish Archive

Europe's most experienced national action has been active since September 1996, when the Swedish National Library initiated the *Kulturarw3* project[5].

The aim of the project is to test methods of collecting, long-term preservation, and providing access to Swedish electronic documents, which are accessible on-line in such a way that they can be regarded as published [MAP00]. Through this project the national library is also laying the foundations of a collection of Swedish electronic publishing. Up to now seven complete sweeps of the Swedish web space have been performed. The Swedish web space does not only contain material from Sweden's domain *.se*, but also *.com*, *.nu*, and numerous others. In total there were more than 60.000 sites harvested only at the last run in spring 2000. The archive currently contains almost 2,5 terabyte of data that is to be stored and managed by a newly purchased tape-robot.

The electronic documents stored include web sites, electronic newspapers, magazines and also - given a somewhat minor priority - ftp-archives and databases. Upon these documents no selection is performed, everything is made as automatic as possible. For this task *NWA-Combine* is used, a modified version of the *Combine Harvesting Robot*[6]. Funded by the European Commission, the crawler was developed by NetLab of Lund University as a part of the *DESIRE* project[7] with the task of harvesting and indexing Internet resources. It, hence, needed to be modified for archiving purposes.

As for preservation of the digital material the aim is to find long-term forms of storage which will facilitate migration to future software and hardware environments. Technically, the archive has been made accessible via a web interface. Thereby it is not only possible to browse the stored documents as if surfing in the Internet at a particular point in time, but also surfing through time, viewing the possibly different versions of the very documents at the times a harvesting run has been performed. An indexing mechanism is planned to complete the archive. As a matter of fact, there is no public access to

[5] `http://kulturarw3.kb.se/html/kulturarw3.eng.html`

[6] `http://www.lub.lu.se/combine`

[7] `http://www.desire.org`

the archive at present because a legal framework is missing. A report of the ministry of education proposes access to the archive should be given to researchers affiliated with recognized institutions. Members of the *Kulturarw3* project, however, see the ultimate aim in securing every citizens right to the free access of information, as it is the case for other legal deposit material.

3.4 Other Projects

To preserve and provide access to the intellectual and cultural heritage the Nordic countries - Denmark, Finland, Sweden, Norway, and Iceland - were amongst the first striving to extend the Legal Deposit Law of each country to electronic media. In 1997 Nordic National Libraries joined efforts on technology development to preserve the web space of the Nordic countries to allow research and public access both today and for the generations to come. The *Nordic Web Archive*[8] started as a forum for coordination and exchange of experience between the different national projects that were in place or about to start. The retrieval and storage is managed and funded by the separate national libraries, but one of the modules, the access module *Nordunet2*, is being developed in shared efforts.

Directed at research on permanent preservation of electronically generated records is the *InterPARES Project*[9]. Located in Canada, it is a major international research initiative in which archival scholars, computer engineering scholars, national archival institutions, and private industry representatives are collaborating to develop the theoretical and methodological knowledge required for the permanent preservation of authentic records created in electronic systems. On the basis of this knowledge it will formulate model strategies, policies, and standards capable of ensuring their preservation [GSE00].

Another very prominent resource on topics related to digital libraries and preservation is the *PADI* web site[10], Australia's *Preserving Access to Digital Information* initiative. It aims at providing mechanisms helping to ensure that information in digital form is managed with appropriate consideration for preservation and future access. As a subject gateway to resources it offers a forum for cross-sectoral cooperation on activities promoting the preservation of access to digital information. An overview of projects addressing aspects of legal deposit of digital publications can be found in [Mui01].

4 The Austrian Pilot Project AOLA

The Austrian On-Line Archive (AOLA)[11] is a cooperation between the Austrian National Library and the Department of Software Technology of the Vienna University of Technology with the goal to make periodical snapshots of the Austrian web space. In an amendment to the Austrian Deposit Law passed in July 2000 off-line electronic media such as CD-ROMs were included in deposit regulations which are to be collected by the Austrian National Libary. Furthermore, as part of this amendment a pilot study addressing the collections and archiving of on-line documents was conceived. Preparations for the pilot-study commenced in 1999, with the first phase having started officially in 2001.

[8] `http://nwa.nb.no`
[9] `http://www.interpares.org`
[10] `http://www.nla.gov.au/padi`
[11] `http://www.ifs.tuwien.ac.at/~aola`

The project is based on a Linux System which, for the pilot phase, is equipped with 240 gigabyte of hard-disk space plus a 6-fold tape-drive for final storage. As for the software, an open-source approach is pursued. This is not primarily for budget reasons, but rather to ensure independence from commercial providers and to easily allow insight into the project. Additionally, this offers the possibility for close cooperation between various other projects in this field. The AOLA project follows a harvesting approach, trying to download the Austrian web space. While doing so, it only archives files that are allowed to be indexed by automatic crawlers by their respective authors, strictly obeying robot exclusion files. Furthermore, if site owners object to their site being included in the archive, yet have failed to specify this in the form of a robot exclusion file, they may have their sites removed from the archive on request. Yet, while we have been in direct contact with numerous site managers who noticed the activity of our crawler, none of them objected to their data being included in the archive so far. We rather experienced high support for our activities, showing the recognition of the importance of such an archive by information providers on the web, and their interest in being part of the Austrian On-Line Archive.

Initially, the project started with the *Nedlib* harvester. Several modifications and expansions had to be made in order to make it fit the needs. Between May 7th and 16th 2001 a first attempt was made to take a snapshot of the Austrian web space. Approximately 1 gigabyte of data was collected per day from the *.at*-domain, as well as several manually selected sites from other domains, amongst them *.com*, *.cc*, *.org* and others. In that time about 666.000 unique URLs were harvested from 1.210 different sites. All in all 8,3 gigabyte of data were stored. During this first pilot run numerous problems with the *Nedlib* harvester were discovered. Some of them could be fixed, as for example the handling of malformed links was updated. Other bugs and errors were too severe to be fixed. After several providers complained that some of their URLs were downloaded again and again blocking other data-traffic, this initial crawl had to be stopped. This multiple downloading of identical sites also was the reason for the rather low download rates of only 1GB per day, with actual data transfer rates being much higher.

Nevertheless, a considerable amount of experience was gained for the second pilot run. The initial run showed that the *Nedlib* harvester basically is constructed such that the requirements for the AOLA project could be met, but for the time being it is not stable enough and it, hence, still needs refinement. As it cannot be foreseen how long it will take until the crawler is available in a stable version, the *Combine* harvester used by the Swedish initiative is now being used. As the *Combine* harvester was initially developed for indexing purposes, rather than for web archiving, several adaptations had to be performed. In doing so the project follows a close cooperation with the Swedish *Kulturaw3* project, benefiting from their experience and efforts already put into modifying the original harvester. Even though, some functionality desirable for an archiving system could not be included so far.

The AOLA project by now has successfully reached the second pilot phase, using the adapted *Combine* harvester. The harvester is currently collecting data at a rate of about 7GB per day, having created an archive of 150 GB, including more than 2,7 million pages from about 21.000 sites by June 21st, 2001.

Apart from problems concerning the collection of the digital material, other aspects mentioned previously have to be considered. To guarantee long-term preservation migration seems to be the only practically applicable approach instantaneously, but for the future a combination with the emulation approach is essential. Furthermore, the technical

aspects to provide access have to be dealt with, requiring a appropriate legal framework. Yet, we again would like to stress the large public support for the creation of such an archive that we have experienced. The importance of these issues is only starting to be recognized to their full extent, with a national strategic paper demanding awareness of the forthcoming "digital culture" [AB00].

5 Conclusions

Early printed books decayed into unrecognizable shreds. Many of the oldest cinematic films were recycled for their silver content [Kah97]. Yet, initiatives to archive the Internet show that humanity is indeed capable to learn from its past. As Brewster Kahle notes, with the Internet "we are locked in the perpetual presence" [Kah01]. While the early days of the Internet have already been lost, there are chances that our cultural heritage that is born digital may do better, allowing us to access our past, to travel in the cyberspace of earlier times, making time-travel a possibility at last.

Yet, serious technical challenges remain to be tackled in order to have our digital creations available even in the near future. Simply collecting them is the first, non-trivial step in this direction, creating large archives of the web. We furthermore have to make sure that these archives are being cared for similar to conventional library archives, maintaining the necessary infrastructure and the physical conditions for keeping the digital objects in good shape. Apart from acquiring these objects, access to them needs to be preserved. As Jeff Rothenberg puts it, "digital information lasts forever – or five years, whichever comes first" [Rot95]. Unless appropriate measures are taken, and unless technical solutions are developed, our archives themselves might end up being no more than vast, unreadable amounts of data. Last, but not least, access to these collections, putting the data to use, requires both legal as well as technical solutions to make these extensive collections navigable, to allow them to be explored.

With the Austrian On-Line Archive (AOLA) project we are taking the first important step towards the creation of such an archive. A first snapshot of the Austrian web space is currently being created, forming the foundation for a comprehensive archive of our digital cultural heritage, capturing the social and economic dimensions as well as the knowledge available in the Austrian web space. Based on this foundation, preservation strategies will allow this content to be available for future generations, preventing it from being "written on the wind" [Bra98].

References

[AB00] H.P. Axmann and H. Badura, editors. *Nationaler Aktionsplan für Österreich (in German)*. Bundesministerium für Bildung, Wissenschaft und Kultur, Wien, Austria, 2000.

[Ban00] S. Band. *The Clock of the Long Now*. Phoenix Publishers, London, UK, 2000.

[Bog96] J.W.C. van Bogart. Media stability studies. Technical report, National Media Laboratory, Final Report, December 1996. `http://www.nml.org/`.

[Bra98] S. Brand. Written on the wind. *Civilization Magazine*, November 1998. `http://www.longnow.org/10klibrary/library.htm`.

[CCS01] CCSDS. CCSDS 650.0-r-1.2: Reference model for an open archival information system (OAIS). Red Book 1.2, Consultative Committee for Space Data Systems, NASA, CCSDS, Mountain View, CA, June 14 2001. http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html.

[CGM01a] B. Cooper and H. Garcia-Molin. Creating trading networks of digital archives. In E. Fox and C. Borgman, editors, *Proceedings of the First ACM/IEEE Joint conference on Digital libraries (JCDL'01)*, pages 353–362, Roanoke, VA, June 24-28 2001. ACM. http://www.acm.org/dl.

[CGM01b] A. Crespo and H. Garcia-Molin. Cost-driven design for archival repositories. In E. Fox and C. Borgman, editors, *Proceedings of the First ACM/IEEE Joint conference on Digital libraries (JCDL'01)*, pages 363–372, Roanoke, VA, June 24-28 2001. ACM. http://www.acm.org/dl.

[Fee99] M. Feeney, editor. *The Digital Culture: Maximising the Nation's Investment.* The British Library, National Preservation Office at the British Library Freepost, London, 1999. http://www.ukoln.ac.uk/services/elib/papers/other/jisc-npo-dig/intro.ht%ml.

[Gra00] S. Granger. Emulation as a digital preservation strategy. *D-Lib Magazine*, 6(10), October 2000. http://www.dlib.org/dlib/october00/granger/10granger.html.

[GSE00] A.J. Gilliland-Swetland and P.B. Eppard. Preserving the authenticity of contingent digital objects: The InterPARES project. *D-Lib Magazine*, 6(7/8), July-August 2000. http://www.dlib.org/dlib/july00/eppard/07eppard.html.

[Kah97] B. Kahle. Preserving the internet. *Scientific American*, March 1997. http://www.sciam.com/0397issue/0397kahle.html.

[Kah01] B. Kahle. Public access to digital materials, Keynote address. In *First ACM/IEEE Joint conference on Digital libraries (JCDL'01)*, Roanoke, VA, June 24-28 2001. ACM.

[Lan96] B. Lang. The legal deposit of electronic publications. Working Series of the General Information Programme and UNISIST CII-96/WS/10, Working Group of the Conference of Directors of National Libraries (CDNL), December 1996. http://www.unesco.org/webworld/memory/legaldep.htm.

[LK98] P. Lyman and B. Kahle. Archiving digital cultural artifacts: Organizing an agenda for action. *D-Lib Magazine*, 4, July-August 1998. http://www.dlib.org/dlib/july98/07lyman.html.

[Lor01] R.A. Lorie. Long term preservation of digital information. In E. Fox and C. Borgman, editors, *Proceedings of the First ACM/IEEE Joint conference on Digital libraries (JCDL'01)*, pages 346–352, Roanoke, VA, June 24-28 2001. ACM. http://www.acm.org/dl.

[MA96] J. Michalko, J.and Garret and D. Aters. Preserving digital information: Final report and recommendations. Technical report, Commission on Preservation

and Access, Task Force on Archiving of Digital Information, May 1 1996. http://www.rlg.org/ArchTF/.

[Man00] J. Mannerheim. The WWW and our digital heritage - The new preservation tasks of the library community. In *66th IFLA General Conference*, Jerusalem, August 2000. IFLA - International Federation of Library Associations and Institutions. http://www.ifla.org/IV/ifla66/papers/158-157e.htm.

[MAP00] J. Mannerheim, A. Arvidson, and K. Persson. The kulturarw3 project - The royal swedish web archiw3e - An example of "complete" collection of web pages. In *66th IFLA General Conference*, Jerusalem, August 2000. IFLA - International Federation of Library Associations and Institutions. http://www.ifla.org/IV/ifla66/papers/154-157e.htm.

[Mui01] A. Muir. Legal deposit of digital publications: A review of research and development activity. In E. Fox and C. Borgman, editors, *Proceedings of the First ACM/IEEE Joint conference on Digital libraries (JCDL'01)*, pages 165–173, Roanoke, VA, June 24-28 2001. ACM. http://www.acm.org/dl.

[Rot95] J. Rothenberg. Ensuring the longevity of digital documents. *Scientific American*, pages 42–47, January 1995.

[Rot99] J. Rothenberg. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*. Council on Library and Information Resources, January 1999. http://www.clir.org/pubs/reports/rothenberg/contents.html.

[WD99] T. van Werf-Davelaar. Long-term preservation of electronic publications: The NEDLIB project. *D-Lib Magazine*, 5(9), September 1999. http://www.dlib.org/dlib/september99/vanderwerf/09vanderwerf.html.