

WebSave - Archiving the Web for Scholar Work

Christian Gütl

Institute for Information processing and Computer supported new Media (IICM),
Graz University of Technology, Austria, (cguetl@iicm.edu)

Infodelio Information Discovery, Co-Initiator of the xFIND Project and Member of the ACM, (cguetl@acm.org)

Wilfried Lackner

Institute for Information processing and Computer supported new Media (IICM),
Graz University of Technology, Austria, (wlackner@iicm.edu)

Abstract: The WWW provides a variety of scholar information, like theses, studies, etc, which may be valuable basics for further and new scholarly work. Referred information of the WWW has to be archived (at the time of writing) for further usage because Internet resources could be changed. The working prototype *WebSave* - described in this paper - support users in their research process. The advanced and novel system manages relevant information of the WWW by importing bookmark files from common browser clients. The system allows adding a wide range of local metadata, storing the referred resources and generates a literature list. Furthermore, it is capable of interacting with other services like preparation for building a background library on CD ROM, building a document archive using *Hyperwave*, building a scholar dynamic Web library by quality metadata exchange to *xFIND*.

Introduction

More and more information, citations and conclusions in scholar works refer to information on the World Wide Web (WWW). This statement may be confirmed e.g. by observations of scholar works and theses at the Institute for Information processing and Computer supported new Media (IICM) at Graz University of Technology. This applies also to publications and reports of conferences and journals. As a matter of fact the WWW is a large – most likely the largest - information repository. However, there is also the serious problem of changing and moving documents on the Internet. It is well documented that the WWW is a highly dynamic Information repository and a perfect media for a fast publishing process. To get revisable Internet references it is necessary to archive and provide the referred information of the WWW for further usage. This may be backed by different basic studies (archiving information from the Internet) by organisations like National Libraries and further institutions. The consideration results in specific requirements within the archiving process of Internet resources for scientific papers and scholarly works. Therefore, the authors of this paper emphasise that it is necessary to suggest possibilities on assistance for users and institutions within their scholar work to archive cited information from the WWW.

The structure of the remaining chapters is given as follows. The first part of the paper, a survey, discusses in brief the archiving process in a historic view as well as the archiving process of electronic documents in general. The second part describes the prototype implementation of the advanced and novel WebSave tool to archive and administrate Internet resources within the process of scholar works, which is based on the conclusions on the first part of the paper.

The Archiving Process of Web Resources

Basically, the motivation for archiving Web pages is to represent snapshots of society for later retrieval and to preserve knowledge for mankind. Therefore, archiving is a long-term process that never ends. Today, search robots or offline browsing tools are used to collect Web sites and to build local archives. However, such tools do not support the process and the requirements of scholarly work. New interactive systems (e.g. chat facilities) require new efforts to gather such documents. It is quite obvious that the archiving process of objects in history is quite different from the requirements on archiving Web pages today, which is discussed as follows.

Archiving from a Historical Perspective

In history, since about 2500 B. C. archives can be found. *Archiving* is the process of collecting and managing objects with historical background. In general these documents are not prepared for publication (e.g. official documents) and are called dead documents (Weiß 2000), because these documents or objects never change. The term *archiving* describes also the classification of documents by formal methods (e.g. grouping, indexing, etc.). Indexing parameters can be subjects, locations and authors. Result of classification tends to an index summary in a printed and / or electronic version.

Three different types of the archiving process can be identified: *Archives* stores only important records or documents, which are not publicized but worthwhile to preserve for mankind; *Libraries* archive published (reviewed) documents; *Museums* archive artistic and scientific documents, images, objects, etc. The storage media and the process of archiving are basically in the analogue sphere (Alscher 1999).

Today, the motivation for archiving is the same, but new technologies (for objects to be archived and within the archiving process) have to be regarded. It is to be mentioned that e.g. on the Internet nearly everyone can publish without any quality control, but also huge valuable amounts of research work and technical reports can be discovered for further scholarly work. A deeper discussion of electronic archiving and detected problems are discussed in the following sections.

Electronic Archiving of Information

The term *electronic archiving* (Weiß 2000) is similar to the concept of imaging (the process of digitizing, storage in databases and retrieval). In general, electronic archiving may be the first step to a digital office. Considering the electronic archiving process, it is subdivided into data storage, data retrieval, and data migration. The process of archiving depends on the characteristics of the document (VSA 2000), which can be categorized in different ways: readability - human and / or machine-readable; type of storage media; document type (text, images, videos, music, etc.).

As a special scope, electronic archiving or imaging can also be understood as the preparatory operations to build Web pages and Web archives. That means the Internet can be used to archive valuable information and enables access also by the Internet. On the other hand, information available on the Internet should be archived, which can also be performed by an Internet archive.

It is well documented that the Internet is a huge - maybe the biggest - knowledge repository of mankind. The Internet offers a cheap and simple way to publish documents without nearly any restrictions. Because of this, the Internet is a disordered, decentralized and mostly not censored worldwide library. It is well known that humans have produced more information since 1945 as in the whole history before. The increase of knowledge is estimated to be exponential. Therefore Internet archives also have to process increasing amounts of data.

Following the idea of archiving (process of collecting and managing), different types of already existing Web archiving methods can be identified like: (1) link collections (e.g. annotated link lists), (2) search engines (e.g. full-text index and document cache), (3) data collections (e.g. ftp) and finally (4) Web archives. All these types provide similar viewing techniques and hierarchical data structures. One advantage of Web archiving methods is the ability of systematic access, which can be performed simultaneously and from different places. Considering the development of mobile communication, this will be increasingly important. The storage is cost effective in relation to the quantity of data. The information can be requested in *real time*. However, one of the main disadvantages is the dynamics and malleability of Internet resources. In addition, information depends on the context, which means that a bit sequence can be interpreted in different ways (digital codes). It is absolutely necessary to use a tool (computer, PDA, etc.) reading and understanding a digital code. Thus, the representation the information depends on the used hardware and software. Because of changing technology over the years, data must be copied to other storage media keeping the information safe (migration). Furthermore, today common transformation rates cause problems related to media objects (video, sound, etc.) performed by remote requests. To sum up, the demands on future archiving systems consist of standardized retrieval interfaces, standardized storage formats and specially standardized document formats. Detected Problems, as followed, has to be considered.

Detected Problems

Some problems related to electronic archiving (focused on Web archives) can be identified as follows:

Information overflow and duplicates: The digitizing of pictures, sound and music data, videos and films produce a huge amount of information. Most of this information should be archived at least for years or even forever. There exist many copies of the same documents at different places.

Information malleability: Information of electronic documents, especially Web resources, could be highly dynamic, and location of documents could be changed or even be deleted. E.g. the average live span of a Web page is about 75 days (Kahle 1996).

Migration and Lifetime of Documents: The capacity of storage media has been increased in the last ten years about hundred times. The transfer rate increases only four times in the same period (Phelps et al. 2000). The result is that the entire set of data cannot be copied in acceptable time. Migration is the process of storing documents permanently. This includes periodical copying into new system environments. Thus, metadata has to be transferred and in general the documents must be converted into a new document format. Parts of information may be lost (VSA 2000).

Access Rights: When people buy a printed journal, they never lose the access right and the content cannot be changed. Online journals can change ownership and therefore the copyright could be changed and the access right will be lost. No one knows how long and who should manage user data concerning access rights.

Related Work and Concluded Requirements

The following paragraphs are focusing especially on archiving Internet resources. Few organizations can be identified, which archive country-specific information. The mentors of these projects are national libraries and governments. They define specific guidelines about registration and documents that have to be archived. Furthermore, single initiatives of public or commercial organizations can also be identified, which effort to archive the entire Internet. A selection of examples is discussed as follows.

EPPP (Electronic Publication Pilot Project) is a project of the National Library of Canada. It has been founded in the time of June 1994 to July 1995. They decided to use the WWW as the primary gateway to access electronic publications. Collection guidelines define that archived documents must be published in Canada, sponsored or produced by a Canadian company (see EPPP 1996).

The *PANDORA* project of the National Library of Australia (NLA) may provide long term access to significant Australian online publications for national preservation. The usage of metadata based on Dublin Core attributes improves the searching process. The online publications are categorized in monographs, serials, home pages and ephemera. The NLA cooperates with other Australian libraries to ensure that there is no duplication of archive materials (see NLA 1997).

ETEXT: Paul Southworth founded the ETEXT archives in the summer of 1992. The project was started in response to the lack of organized archiving of political documents and discussions disseminated via Usenet on newsgroups. In the last five years three GB of mostly ASCII text are stored (see <http://www.etext.org>).

PURL: The PURL (Persistent URL) service provides an archive server solution avoiding the "404 page not found" error. A PURL is like an URL and associates the PURL with the actual URL and returns that URL to the client. The client can then complete the URL transaction in the normal way. PURL's are the result of OCLC's URN (Uniform Resource Name) standard and library cataloguing communities (see Shafer et al. 1996).

The *Internet Archive (IA)* is an association, which collects and stores materials from the public Internet (WWW, Gopher, Newsgroups, FTP). Any Web page that uses CGI requests or needs to authenticate to get access is blocked by the gathering procedure. The archive is not yet publicly available but provides free access to researchers, historians and scholars. The IA archive size is about 14 TB (mostly Web sites) since 1996 to the time of writing this paper (see <http://www.archive.org>).

Alexa: Alexa is a free Web Navigation Service that gives a public access to the IA and rates Web sites of IA. The navigation tool works with common Web browsers. Information about the page recently visited are available (related links, quality, traffic, actuality, etc.). Alexa provides a "non-dead link" service by offering archived pages of IA (see <http://www.alexa.com>).

Google: Google is a fast search engine which uses the patented PageRankTM – technology. PageRank counts incoming links to a document and assesses also outgoing links. "Google stores many web pages in its cache to retrieve for you as a back-up in case the page's server temporarily fails" (see <http://www.google.com>).

Offline browser: Offline browser tools gather single HTML pages or structures including embedded objects (images, scripts, audio and video files). Such systems follow the idea that reading takes much more time than to replicate the pages. Thus, the reading process can be done offline. Examples of free tools are WebStripper™ (see <http://webstripper.net>) and WebCopier (see <http://www.maximumsoft.com>.)

Based on the survey (see above) and the related work (stated so far), the authors of this paper propose a solution in the field of scholarly work. Existing commercial and non-commercial mirroring tools provide only offline browsing on the local file system. Furthermore, online archiving systems cannot guarantee that documents of interest are archived. Therefore, the proposed scholarly solution follows a different idea. Users have to be supported in the process of collecting relevant information from the WWW or Internet by less or even no further effort. In addition, they have to be supported to manage these resources and enabled to enrich these resources with remarks, quality metadata, etc. Because of the human effort within the process for discovering *good* and *relevant* information, such resources and its meta-information should also be provided other users, research teams, scholar organisations and Internet communities for further research. Based on the results stated so far, the *WebSave tool* was designed, which is discussed in the following Chapter.

The WebSave Tool

On a glance, the *WebSave* tool is a Java based prototype implementation, which supports users at their research process within the scholar sphere. Unlike existing offline browsing tools, the advanced and novel system manages relevant information of the WWW by importing bookmark files from common browser clients. First, the users' view of the *WebSave* tool is discussed, followed by brief information of the administrators' view and the cooperation with the xFIND system. A more detailed discussion of the archiving of resources in electronic media and a detailed description of the *WebSave* tool can be found in Lackner (2000).

Simply by using a common Web client, interesting links can easily be collected as *bookmarks* placed in particular folders. Different folders can be assigned to different projects or chapters of scholarly works. The *WebSave* tool imports these *link collections* and supports the user to manage these resources. In addition, any of these projects can be titled and described in an abstract by additional attributes. It should be noted that also incremental updates from the bookmark list could be performed. That means that users can search for relevant information on the WWW, set bookmarks and finally import the bookmark information in the *WebSave* environment. Within another session, further relevant information can be added to the corresponding bookmark folder and incremental imported. Because of the short life of Web sites, it is recommended to gather the documents of interest within some days (at the time of writing a paper).

WebSave append additional information (*WebSave* system information and metadata) to the imported bookmarks. System information is e.g. identifier, last visit time, last modification time, etc. Metadata are e.g. notations as well as quality metadata (xQMS), etc. The xQMS (xFIND quality metadata scheme) consists of descriptive (e.g. subject domain, etc.) and evaluative (e.g. authority, target audience, etc.) quality metadata. xQMS is developed by the xFIND group (<http://xfind.iicm.edu>) and described in detail in Weitzer (2000). In addition, the *WebSave* tool provides literature entries of the resources by compiling reference identifiers, the reference descriptions (author, organisation, year of publication, title, version), the time of last visit and finally the time of the saving process (see also below).

Furthermore, the resources can be gathered and locally stored on the file system. That means that a relevant bookmark imported by *WebSave* refers to a document available as HTML with embedded graphics, scripts, applets, etc. The tool stores the HTML files as well as the embedded objects. The *WebSave* tool manages several projects for any user and allows the cooperation of users. After finishing a project the literature list can be generated automatically by the system. Furthermore, *WebSave* generates an HTML version of the list of literature and prepares links to the local (cached) Internet resources. *WebSave* supports the preparation of the Internet resources and its metadata for a CD-ROM production. The HTML reproductions of a project comprise the project abstract, summary, reference list including links to local stored sources as well as links to the original Internet sources. Users who are finishing diploma theses can also provide an electronic version of their work and in addition provide local (saved) versions of referred Internet resources (CD-ROM version of the work).

WebSave allows the usage by more than one user. All data are stored in a subfolder of a predefined network directory. The subfolder contains the entire set of user data (personal data, project data, etc.), which can be managed by the user and administrators. Within any subfolder, Web resources are stored also project dependent and managed in the original hierarchy. Also the embedded objects of HTML files are mirrored within

the same structure if these objects are located on the same origin then the HTML files. It is to be mentioned that duplications of objects are enabled, because in different projects different version of objects can occur.

The WebSave administrator view provides more options. Administrators can admin a group of user (e.g. a course, a research group, a research institution, etc.) by adding, editing and removing user records. Administrators are enabled to collect user folders or user projects building a scholarly knowledge pool. An HTML reproduction of the knowledge pool can be built to place them at the disposal of an institution library (e.g. CD-ROM, Intranet, etc.). They also allowed to evaluate, edit and add xQMS metadata (Fig. 1). In addition, they can prepare the repository to render them searchable by the xFIND search system (see below).

xQMS help		Predefined xQMS meta data set:
Scope Of Description	2	
Type	13	
Topic	02	
Alternative Topic	H.3	
Classification Scheme	ACM	
Class Scheme Ident	http://www.acm.org/class/1998/overview.html	
Keywords	Internet Archiving, Storing Technologies, Preservation of Digital History	
Description	Describes the technical possibilities of archiving Web sites in consideration of	
Quotations Hints	Kahle, B.: Archiving the Internet; 11.4.1999, last visit Nov. 2000	
Language	en-us	

Figure 1. Web resource metadata editing form

As already stated above, discovering relevant information causes substantial human effort (a scientist looks up to 100 documents before finding a relevant one). Because of that, *good* and *relevant* information as well as additional information (descriptive and evaluative quality metadata) should also be archived and provided for further scholarly works by users, organisations and communities. In addition, these (local) metadata have to be used to improve the retrieval process of the resources. A possible solution meeting this requirement is the novel xFIND (extended Framework for Information Discovery) system (see xFIND 2000), which can build an index of the local Internet resources. Furthermore, the system combines the full-text index of the resources with the quality metadata and allows an improved search process (e.g. looking for a peer reviewed paper containing the phrase “*web based training*” in the full-text). In addition, the system also allows to index the original sources and to track changes of the resources. In a further version, a notification (e.g. user who referred the Web resource) will be performed. The human-created metadata (provided by the WebSave tool) and the computer-automated information (processed by the xFIND system) of the resources represent a helpful and growing information repository based of references of scholarly works. The idea follows the strategy of causing little effort by any individual within their research process (e.g. working on a paper), and accumulated output (a big background library of distributed Web resources and valuable metadata for the retrieval process) for the scholar communities or mankind. It is worth mentioning that the aim of the xFIND system is to build up a distributed, huge knowledge management system for Web resources enriched with descriptive and evaluative metadata.

Current and Future Work

Current work is going on to develop an interface layer for data exchange between the xFIND system and the WebSave tool. As a future work, in addition to local archiving of Web resources, an interface for an online document archiving system will be implemented. The first prototype implementation will be done on a Hyperwave Information Server (see <http://www.hyperwave.com>), which supports versioning and an advanced rights management. Thus, new relevant information of Web resources provided by the WebSave tool is transmitted to the xFIND system. The later system gathers the new documents and renders the full-text as well as the additional metadata searchable. In addition a document repository is requested to archive the Web resource. If the xFIND system detect any changes of the document, the document archiving system will be informed to archive also the new version.

Conclusion

It is obvious that increasingly Web resources are cited in scholar works. However, the problem of changing and removing of these resources requires solutions to preserve this referred information for further access. The WebSave tool in combination to the xFIND system is a possible solution to counteract the shortcomings stated so far. First experiences by the usage of students have shown that the tool will support their scholarly work preparing studies and diploma theses. The preservation of valuable Web resources and the enrichment of quality metadata allow to build up a growing scholarly knowledge repository, which supports further scholarly work. The xFIND system can be perfectly used to renders such repository searchable and provides an interface to the original sources and tracks changes of them.

References

- Alscher, H. J. (1999). Organisation und Geschichte des Bibliothekswesen; May 1999, last visit Nov. 2000
<http://sites.netscape.net/hansjoachimalscher/BIBLKURS/biblog.htm>
- EPPP (1996). Electronic Publications Pilot Project, Summary of the Final Report, National Library of Canada, 07.05.96, last visit Nov. 2000. <http://www.nlc-bnc.ca/pubs/abs/eppp/esumreport.htm>
- Kahle, B. (1996). Archiving the Internet; 11. Apr. 1999, last visit Nov. 2000. http://www.archive.org/sciam_article.html
- Lackner, W. (2000). Archiving the Web for Scholar Work, IICM TU-Graz Austria, Okt. 2000, last visit Nov. 2000.
<http://www2.iicm.edu/cguetl/education/projects/WebSave>
- NLA (1997). National library of Australia: Jasmine Cameron. PANDORA - Preserving and Accessing Networked DOcumentary Resources of Australia, Review of progress to June 1997; 7. Dec. 1999, last visit July 2000.
<http://www.nla.gov.au/policy/pandje97.html>
- Phelps, A., T. & Wilensky R. (2000). Division of Computer Science; University of California, Berkeley, Robust Hyperlinks Cost Just Five Words Each, UCB Computer Science Technical Report UCB//CSD-00-1091. 10. Jan. 2000, last visit Nov. 2000. <http://http.cs.berkeley.edu/~wilensky/robust-hyperlinks.html>
- Shafer, K. & Weibel, S. & Jul, E. & Fausey, J. (1996). OCLC, Online Computer Library Center, Introduction to Persistent Uniform Resource Locators; last visit Nov. 2000. http://www.isoc.org/inet96/proceedings/a4/a4_1.htm
- VSA (1999). Verein Schweizerischer Archivarinnen und Archivare: Archivieren im Informationszeitalter; last printed version: 1.0, 31. Mar. 1999, online version: 1.0.4, 26. Apr. 1999, last visit Nov. 2000.
http://www.staluzern.ch/vsa/ag_aea/dok/Basisdokument_d.html
- Weiß D. (2000). Informationen über elektronische Archivierung. Archiv & Workflow; 26. Mar. 2000, last visit Nov. 2000.
<http://www.dr-weiss.com/archiv01.htm>
- Weitzer J. (2000). Verwendung von Qualitäts-Metadaten zur verbesserten Wissensauffindung und Testimplementierung im xFIND System, IICM TU-Graz Austria, 31. Mar. 2000, last visit Nov. 2000.
<http://www2.iicm.edu/cguetl/education/thesis/jweitzer>
- XFIND (2000). The Official xFIND Homepage. IICM TU-Graz Austria, last visit Nov. 2000. <http://xfind.iicm.edu>

Acknowledgements

We thank all of members of the IICM and the xFIND team, who have supported the work on the WebSave tool. Especially, we thank Prof. Maurer, who has inspired our work.

Furthermore, we thank Infodelio (<http://www.infodelio.com>), which has supported us finishing this paper and has financed the conference fee and travel costs. Infodelio is working and researching e.g. in the field of Web resource discovery for research departments and scholar institutes.