

## D-Lib Magazine December 2002

Volume 8 Number 12

ISSN 1082-9873

# Uncovering Information Hidden in Web Archives

## A Glimpse at Web Analysis building on Data Warehouses

Andreas Rauber

<andi@ifs.tuwien.ac.at>

Andreas Aschenbrenner

<andreas.aschenbrenner@nationaalarchief.nl>

Oliver Witvoet

<a9503280@unet.univie.ac.at>

Robert M. Bruckner

<bruckner@ifs.tuwien.ac.at>

Max Kaiser

<max.kaiser@onb.ac.at>

---

### Abstract

The Internet has turned into an important aspect of our information infrastructure and society, with the Web forming a part of our cultural heritage. Several initiatives thus set out to preserve it for the future. The resulting Web archives are by no means only a collection of historic Web pages. They hold a wealth of information that waits to be exploited, information that may be substantial to a variety of disciplines. With the time-line and metadata available in such a Web archive, additional analyzes that go beyond mere information exploration become possible.

In the context of the *Austrian On-Line Archive (AOLA)*, we established a Data Warehouse as a key to this information. The Data Warehouse makes it possible to analyze a variety of characteristics of the Web in a flexible and interactive manner using on-line analytical processing (OLAP) techniques. Specifically, technological aspects such as operating systems and Web servers used, the variety of file types, forms or scripting languages encountered, as well as the link structure within domains, may be used to infer characteristics of technology maturation and impact or community structures.

### Introduction

In recent years, we have seen not only an incredible growth of the amount of information available on the Web, but also a shift of the Web from a platform for distributing information among IT-related persons to a general platform for communication and data exchange at all levels of society. The Web is being used as a source of information and entertainment; forms the basis for e-government and e-commerce; has inspired new forms of art; and serves as a general platform for meeting and communicating with others via various discussion forums. It attracts and involves a broad range of groups in our society, from school children to professionals of various disciplines to seniors, all forming their own unique communities on the Web. This situation gave rise to the recognition of the Web's worthiness of being archived, and the subsequent creation of numerous projects aiming at the creation of World Wide Web archives. Snapshot-like copies of the Web preserve an impression of what hyperspace looked like at a given point in time, what kind of information, issues, and problems people from all kinds of cultural and sociological backgrounds were interested in, the means they used to communicate their interests over the Web, characteristic styles of how Web sites were designed to attract visitors, and many other facets of this medium.

Several initiatives are already building bulk-collection Web archives, with the two longest-running initiatives being the US based *Internet Archive* [13, 14], which currently holds the largest collection of online material from all over the world, and the *Kulturaw3* project at the Royal Library in Sweden [[4, 25], containing an archive of frequent snapshots of the Swedish Web since 1996. Similar pilot projects exist at the national libraries of France, Iceland, the Czech Republic, Estonia, and other nations. A major initiative integrating Web harvesting into a generic digital deposit framework is the *NEDLIB* project [10, 20].

Alongside those initiatives employing Web crawlers to extract a comprehensive impression of the Web, selective approaches attempt to identify and filter relevant assets manually. The *PANDORA* project [19] at the National Library of Australia has pioneered this policy. Numerous institutions follow similar objectives, such as the national libraries of Tasmania, Germany, and the Netherlands, as well as the British Library.

When it comes to archive usage (or prospected usage, as many of these archives currently cannot provide access to their collections due to legal or technical reasons), most projects focus solely on making the collected Web material accessible. Interfaces allow users to surf through time and see the evolution of a Web page from one crawl to the next. The Internet Archive constructed the *Wayback Machine* [13] as an interface to its repository. Recently the *Nordic Web Archive* initiative [21] released the *NWA Toolset* to offer intuitive Web browsing through time.

A significant challenge in this domain is the preservation of these collections over the long term. Though numerous projects address digital preservation, the question of how to maintain Web material so that it is accessible and authentic over the centuries prevails [7, 9, 27]. For pointers to the different initiatives, see the Web pages of the ECDL Workshop on Web Archiving [17], the AOLA bibliography [3], and the *PADI* Website [18] of the National Library of Australia.

With such a repository of Web data, as well as the metadata that is associated with the

documents and domains, we have a powerful source of information that goes beyond the content of Web pages. The Web is not only content, it is rather, technically speaking, a medium transporting content in a variety of ways, using different technical platforms, as well as data representations to make its information available. The providers of information are located in different physical places on the hyper-linked world, and information is transferred via a variety of channels. Having an archive of the World Wide Web means that, not only can we see which information was available at which time, we can also trace which technology was used for representing a certain kind of information and what kinds of systems were used to make the information available. Web archives also give us the means to monitor the life cycle of technology, following file formats, interaction standards, and server technology from their creation, through different degrees of acceptance, to either prolonged utilization or early obsolescence. This knowledge, in turn, may influence decisions for digital preservation. It may also form the basis for technology selection in projects, with both the stability of a given technology and its diffusion being key issues for project success and sustainability.

In order for the most useful analyses to yield answers to project questions and issues, a different perspective of the Web and Web archives is needed, a perspective focusing not solely on content, but on the wealth of information automatically associated with each object on the Web. The information about the object could include its file format, size and currency; the object's link structure and connectivity to other pages within the same site and domain; and, externally, the language used, operating system and Web server software running on the server side machine; the use of specific protocols and cookies; and many more types of information. Among the projects involved in researching methods for Web analysis are:

- *Web Archaeology* [16], which uses a variety of content representations to study how the Web evolves over time.
- *WebBase* [11], which addresses issues such as the functional design, storage management, as well as indexing modules for Web repositories.
- *WHOWEDA* [26], which employs Data Warehouse technology for the storage of consecutive versions of Web pages, adding a time dimension to the analysis of content and link structure.

We also address these issues within the scope of the *Austrian On-Line Archive*, a joint initiative by the Austrian National Library and the Vienna University of Technology, to analyze and devise ways for archiving the Austrian national Web space. We adopt a solution based on a Data Warehouse for the *Austrian On-Line Archive Processing* module (AOLAP) [24], allowing interactive analysis of the accumulated data using on-line analytical processing techniques.

## **Data Warehousing and OLAP**

When it comes to analyzing large amounts of data in a flexible manner, *Data Warehouses* (DWH) have evolved into the core components of Decision Support Systems [15, 22]. A Data Warehouse is a subject-oriented, integrated, time-variant, non-volatile collection of

data in support of decision-making processes [12]. Rather than storing data with respect to a specific application, the information is processed for analytical purposes, allowing it to be viewed from different perspectives in an interactive manner. It furthermore integrates information from a variety of sources, thus enriching the data and broadening the context and value of the information.

The primary concept of a DWH is the separation of information into two main categories, referred to as *facts* and *dimensions*, respectively. *Facts* are the information that is to be analyzed, with respect to its dimensions, which often reflect business perspectives, such as a geographic location, evolution over time, product groups, merchandising campaigns, or stock maintenance. The DWH may be envisioned as a multi-dimensional data cube. This data cube allows us, using *on-line analytical processing* (OLAP) tools, to interactively drill-down, roll-up, slice and dice, view and analyze the data from different perspectives, and to derive ratios and compute measures across many dimensions. These OLAP operations assist in interactive and fast retrieval of 2D and 3D cross-tables and chart-table data from the cube, which allow convenient querying and analysis of a Web data storage.

### **AOLA: The Austrian On-Line Archive**

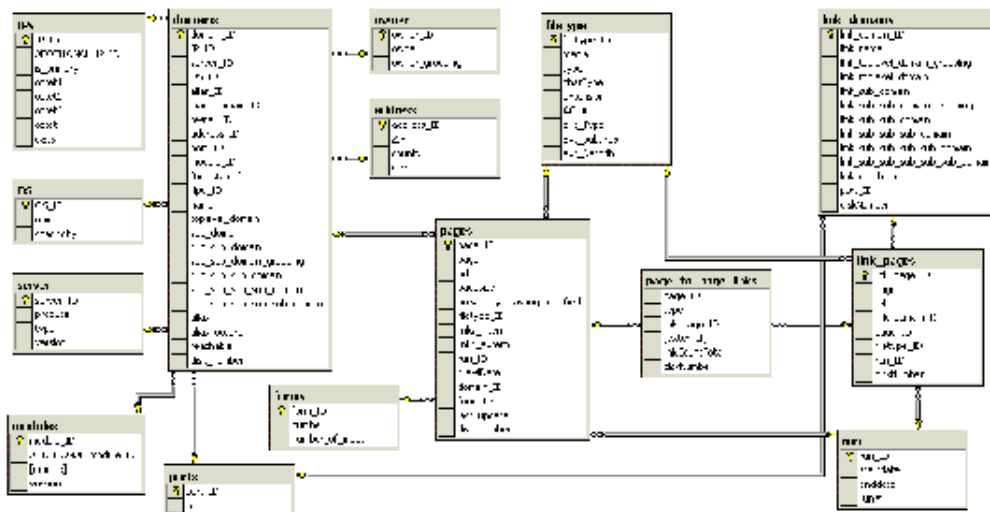
The *Austrian On-Line Archive* (AOLA) [2, 23] project aims at creating a permanent archive documenting the evolution of the Austrian Web space. AOLA is being conducted by the Austrian National Library [1] and the Department of Software Technology and Interactive Systems [8]. The Austrian National Library has been active in the field of Web archiving for several years. AOLA, which started in 2000, uses a bulk collection approach for the Austrian Web space. It complements a pilot project on selective archiving of on-line publications that ran at the Austrian National Library from 1997-1999. In the scope of the AOLA project, the Austrian Web space covers the whole *.at* domain, as well as servers located in Austria yet registered under "foreign" domains like *.com*, *.org*, *.cc*, *etc.* Furthermore, sites dedicated to topics of Austrian interest (so-called "Austriaca") are considered, even if they are physically located in another country. Austrian representations in a foreign country like the Austrian Cultural Institute in New York (at <http://www.acfny.org>) are examples for such sites of interest. The inclusion of these servers, so far, is determined semi-automatically by maintaining a list of allowed non-at servers.

A modified version of the Combine Crawler [5], is used to gather the data from the Web. While the crawling process itself runs completely automatically, manual supervision and intervention is required (e.g., when faulty URLs are encountered). The pages down-loaded from the Web are stored together with automatically acquired metadata in a hierarchical structure and are archived in compressed format on tapes.

The archive currently consists of about 488 GB of data from two partial crawls with more than 2.8 million pages from about 45,000 sites from the first crawl in 2001 (118 GB in total), as well as about 370 GB (approximately 8.2 Mio URLs from about 170,000 servers including alias names) from the second crawl in spring 2002. It contains all types of files as collected by the harvesting software. In addition to the actual pages, metadata obtained automatically during the crawling process is stored as part of the archived files. This includes information provided as part of the *http* protocol as well as other information

provided by the server, including: the server software type and version, the operating system used by the server, date and time settings at the server, and last-modified dates for the respective file being down-loaded.

The information extracted from the pages includes: *file types* based on file extensions and the associated MIME type obtained from the Web server, *file size*, internal and external *links*, information about *e-mail addresses* and interactive *forms* used in the case of *HTML* files, *date of last modification*, and others. Concerning the various domains, we mainly concentrate on *IP addresses* and thus *network types*, *operating system* and *Web server software* information. Furthermore, we integrate information from other sources to enrich the data provided by the harvesting system. Specifically, we use a set of WHOIS servers to provide geographic location information of Web service registrars, alias names, etc.



**Figure 1: Database model for the AOLAP system**  
(For a larger view, click here.)

The information is further transformed and loaded into a relational database using a star-model-like design for the data storage. The data model basically consists of two parts, as depicted in a simplified manner in Figure 1. The first part arises from all tables containing data about the Web hosts from which the data derives. The second part consists of the tables containing data about the hosts to which links point. Connecting these parts is the table where all the links are stored. This table forms the central fact table in the Data Warehouse. Based on these tables, a multi-dimensional cube is created which can further be used for interactive analysis.

## Experimental Results

### Distribution of file-types over different Web servers

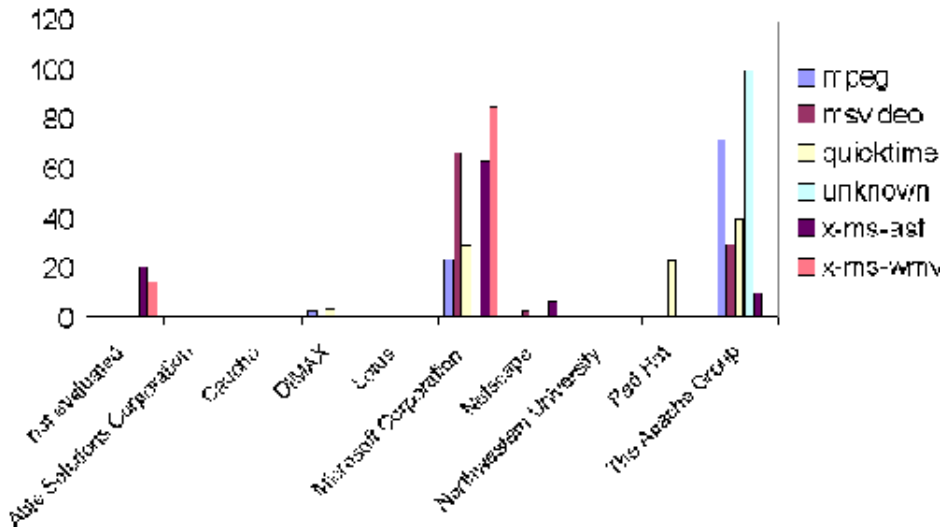
The number of file types encountered in the Web archive is highly relevant with respect to the preservation of the archive, that is, keeping the pages viewable in the near and far future. The number of types also represents a good mirror of the diversity of the Web with respect

to the technologies employed for conveying information. Overall, we encountered more than 200,000 different types of files based on their extensions, and more than 200 different types of information representation when we use the MIME type as the indicative criterion. (However, we should stress that the quality of the information provided this way is rather low, as a large number of both file extensions as well as MIME types are actually invalid, such as files with extensions *.htmo*, *.html* or *.median*, *.documentation*.) A listing of some of the most important types of files found in the archive is provided in [Table 1. For a comprehensive overview of almost 7,000 different file extensions and their associated applications, see *FILExt: The file extension source* [6]. While the majority of file extensions encountered definitely are erroneous, the erroneous extensions are indicators of serious problems with respect to preserving that kind of information, as well as the need to define solutions for cleaning this dimension to obtain correct content type descriptors.

**Table 1:** Selection of MIME types encountered

MIME type	#Occ.	MIME type	#Occ.
Application/ms-excel	1227	Image/gif	35144
Application/ms-powerpoint	841	Image/jpeg	145200
Application/msword	14799	Image/png	349
Application/octet-stream	9916	Image/tiff	1025
Application/pdf	67976	Image/x-bitmap	426
Application/postscript	5274	Image/other	123
Application/x-dvi	634	Text/css	713
Application/x-msdos-program	1231	Text/html	7401473
Application/x-tar	2189	Text/plain	32549
Application/x-zip-compressed	15314	Text/rtf	2783
Application/other	6985	Text/vnd.wap.wml	2961
Audio/basic	246	Text/other	753
Audio/x-mpegurl	3947	Video/mpeg	983
Audio/x-midi	1777	Video/msvideo	596
Audio/x-mpeg3	3240	Video/quicktime	768
Audio/x-pn-realaudio	5006	Video/x-ms-asf	646
Audio/x-wav	1430	Video/unknown	4
Audio/other	671	Video/other	20

Several interesting aspects can be discovered when analyzing the distribution of file types across the different types of Web servers. Generally known tendencies, like the dominance of the *Portable Document Format (PDF)* over the previously very important *Postscript* file format for document exchange, can be verified this way.



**Figure 2: Distribution of video file types across Web servers**

(For a larger view, [click here.](#))

Figure 2 depicts the distribution of various types of video file formats across Web servers. Here we find significant differences in the way video information is provided with respect to the type of Web server employed. *Mpeg* is by far the dominant format on Apache Web servers, followed by *QuickTime*, which is less than half as popular but still ahead of various other video formats identified by their MIME type as flavors of *ms-video*. (We also find a video format identified as MIME type *video/unknown* on Apache servers. By drilling down the associated file extension dimension, these files were identified to be *.swi* and *.raw* files—the former, for example, being a *Swish* data file used in connection with *Flash* animations.)

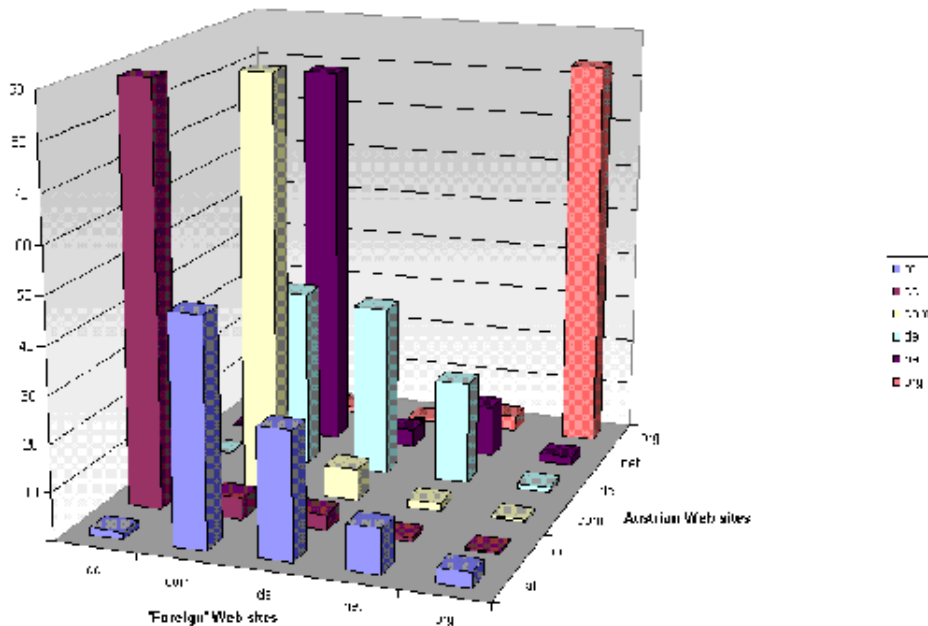
This is in sharp contrast to the situation encountered at Web sites running the MS IIS Web server where the family of *ms-video* and *ms-asf* formats by far dominate the type of video files provided. The *ms-asf* format is used by MS Active Streaming (Media) files containing audio and/or video data and compressed with 3rd party codes. When we take a look at the Netscape Web server, we again find a slight dominance of *ms-video* file formats. Another interesting characteristic is exhibited by the Stronghold Web server (the Red-Hat Secure Web server for Linux operating systems), which—when it comes to video files—provides only *QuickTime* movies, as shown in Figure 3. Untypical distributions like this example may quite frequently be attributed to artifacts such as a single Web server running a specific system and providing a large amount of files as part of a collection. The Data Warehouse allows us to interactively drill-down on this section and reveals, in this case, the distribution can be attributed to a sub-group of 10 domains out of several hundred domains using the Stronghold server. Of these 10 domains, however, 9 are closely related to each other and are part of one larger organization providing identical information, thus actually being a kind of mirror of one site. Due to the flexibility of the interactive analysis facilitated by the DWH, these artifacts can easily be identified.

Ext. Type	Ext. Subtype	Page Count	Page Size Total	Page Size Avg	Page Size Min	Page Size Max	Page Size Std Dev	Page Size Total	Page Size Avg
Q (QuickTime)	Q (QuickTime)	1,773	4,122	2,324	1,024	4,122	1,773	4,122	2,324
+	+	17	1	1	1	1	17	1	1
+	+	1,756	4,121	2,323	1,024	4,121	1,756	4,121	2,323
+	+	11	265	24	1	265	11	265	24
+	+	190	99	52	1	99	190	99	52
+	+	52	91	17	1	91	52	91	17
+	+	9	1	1	1	1	9	1	1
+	+	7,709	1,363	175	1	1,363	7,709	1,363	175
+	+	101	9	9	1	9	101	9	9
+	+	300	1,271	424	1	1,271	300	1,271	424
+	+	1,000	1,000	1,000	1	1,000	1,000	1,000	1,000
+	+	1,966	77,416	39,382	1	77,416	1,966	77,416	39,382
+	+	100	100	100	1	100	100	100	100
+	+	180	180	180	1	180	180	180	180
+	+	1	1	1	1	1	1	1	1

**Figure 3: Stronghold servers providing only QuickTime movies**  
(For a larger view, click here.)

### A glimpse at the link structure

In this section we take a brief look at the link structure within the Austrian Web space and, more specifically, on the link characteristics within the various top-level domains as depicted in Figure 4.



**Figure 4: Link structure within top-level domains in .at**  
(For a larger view, click here.)

A first glance at Figure 4 confirms an intuitive tendency: namely, the fact that we have a

high inter-linkage within each respective domain, i.e., *.com* sites linking mostly to other *.com* sites, *.cc* linking within *.cc* and so on. However, we also find some interesting exceptions to this rule, such as *.de* that while having a significant number of links within its domain, has a higher link count to sites located at *.com*. This characteristic is even more observable for the domain *.net*, which has by far the strongest inter-linkage with *.com*. To analyze the reason for these numbers, we can perform a drill-down on the *.net* domain, as depicted in Figure 5. This drill-down reveals that the majority of these *.net* links are originating from one server within Austria (<tu cows-server.austria.net>). By drilling down the target domains dimension, we furthermore find that most of these links are pointing to the respective *tucows* server in the *.com* domain.

Level	Sublevel	Sublevel	Sublevel	Sublevel	Sublevel	Sublevel	Sublevel	Sublevel	Sublevel	Sublevel	Sublevel	Sublevel	Sublevel	Sublevel	Sublevel	Sublevel	Sublevel	Sublevel	Sublevel	Sublevel	
cc	de	cc	com	de	cc	com	de	cc	com	de	cc	com	de	cc	com	de	cc	com	de	cc	com
Link Index	12345	6789	1011	1213	1415	1617	1819	2021	2223	2425	2627	2829	3031	3233	3435	3637	3839	4041	4243	4445	4647

**Figure 5: Link structure: drill-down on *.net***  
 (For a larger view, click here.)

### Searching for relations between different sites in the Web space

In this section we take a look at communities identifiable in the Austrian Web space within the *.at* domain. In general, we find the same tendency as for the top-level domains, i.e., a high inter-linkage within, for example, the academic, organizational or commercial domain.

During the interactive analysis, we can also take a closer look at some specific domains exhibiting a different characteristic. One example is the web page at <http://www.asn-linz.ac.at>., part of the 'Austrian School Network Linz'. The page is a portal to educational offers in Austria, containing links to all the schools and universities in Austria. When analyzing the outgoing links from this domain, however, we find the dominant domain to be a non-academic site <http://www.geolook.at> (cf. Figure 6). This site provides maps of Austria and is used by the portal as a location indicator for each referenced university or school. Consequently, we find a local hub-authority relation between these two domains.



facilitated by the collection and integration of additional information-such as link structure analysis-far more fascinating insights into the Web and its evolution will become possible. These insights will provide a basis for crucial technology decisions including: the evolution and maturation of technologies, analysis of market shares, and, from a preservation perspective, technologies and efforts required to preserve the diversity of information representation.

Obviously, further types of information can be extracted from the Web pages and integrated into the Data Warehouse (e.g., automatic language detection methods) covering in larger detail additional technological information, such as the usage of cookies, embedded Java applets, Flash plug-ins, encryption, and others. Furthermore, being able to analyze the content-based dimension of a Web archive provides the basis for subject gateways on a variety of topics and with a historic dimension.

## References

- [1] Austrian National Library. Website.  
<<http://www.onb.ac.at>>.
- [2] AOLA. Austrian On-Line Archive. Website.  
<<http://www.ifs.tuwien.ac.at/~aola>>.
- [3] AOLA. Austrian On-Line Archive - bibliography. Website.  
<<http://www.ifs.tuwien.ac.at/~aola/links/WebArchiving.html>>.
- [4] A. Arvidson, K. Persson, and J. Mannerheim. The Kulturarw3 project - The Royal Swedish Web Archiw3e - An example of "complete" collection of web pages. In *Proceedings of the 66th IFLA Council and General Conference*, Jerusalem, Israel, August 13-18 2000.  
<<http://www.ifla.org/IV/ifla66/papers/154-157e.htm>>.
- [5] Combine harvesting robot. Website.  
<<http://www.lub.lu.se/combine>>.
- [6] Computer Knowledge (CKNOW). FILExt: The file extension source. Webpage.  
<<http://filext.com/>>.
- [7] M. Day. Metadata for digital preservation: A review of recent developments. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2001)*, Springer Lecture Notes in Computer Science, Darmstadt, Germany, Sept. 4-8 2001. Springer.  
<<http://link.springer.de/link/service/series/0558/papers/2163/21630161.pdf>>.
- [8] Department of Software Technology and Interactive Systems (IFS), Vienna University of Technology. Website.  
<<http://www.ifs.tuwien.ac.at>>.
- [9] ERPANET: Electronic Ressource Preservation and Access Network. Website.

<<http://www.erpanet.org/>>.

[10] J. Hakala.

Collecting and preserving the web: Developing and testing the NEDLIB harvester. *RLG DigiNews*, 5(2), April 15 2001.

<<http://www.rlg.org/preserv/diginews/diginews5-2.html>>.

[11] J. Hirai, S. Raghavan, H. Garcia-Molina, and A. Paepcke. Webbase: A repository of web pages. In *Proceedings of the 9th International World Wide Web Conference (WWW9)*, Amsterdam, The Netherlands, May 15-19 2000. Elsevier Science.

<<http://www9.org/w9cdrom/296/296.html>>.

[12] W. Inmon. *Building the Data Warehouse*. J.Wiley and Sons, New York, NY, 1992.

[13] Internet Archive. Website.

<<http://www.archive.org>>.

[14] B. Kahle. The Internet Archive - editor's interview with Brewster Kahle. *RLG DigiNews*, 6(3), June 15 2002.

<<http://www.rlg.org/preserv/diginews/diginews6-3.html>>.

[15] R. Kimball. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. John Wiley & Sons, 2 edition, 2002.

[16] S. Leung, S. Perl, R. Stata, and J. Wiener. *Towards web-scale web archeology*. Research Report 174, Compaq Systems Research Center, Palo Alto, CA, September 10 2001.

<<http://gatekeeper.dec.com/pub/DEC/SRC/research-reports/SRC-174.pdf>>.

[17] C. Lupovici, J. Masanes, and A. Rauber. Second ECDL Workshop on Web Archiving, September 19 2002. Website.

<<http://bibnum.bnf.fr/ecdl/2002/index.html>>.

[18] National Library of Australia. PADI: Preserving Access to Digital Information. Website.

<<http://www.nla.gov.au/padi>>.

[19] National Library of Australia. PANDORA: Preserving and Accessing Networked Documentary Resources of Australia. Website.

<<http://pandora.nla.gov.au>>.

[20] NEDLIB project. Website.

<<http://www.kb.nl/coop/nedlib>>.

[21] Nordic Web Archive. Website.

<<http://nwa.nb.no>>.

[22] T. Pedersen and C. Jensen. Multidimensional database technology. *IEEE Computer*,

34(12):40-46, December 2001.

[23] A. Rauber and A. Aschenbrenner. Part of our culture is born digital - On efforts to preserve it for future generations. *TRANS. Internet-Zeitschrift für Kulturwissenschaften (Internet Journal for Cultural Studies)*, 10, July 2001.  
<<http://www.inst.at/trans/10Nr/inhalt10.htm>>.

[24] A. Rauber, A. Aschenbrenner, and O. Witvoet. Austrian on-line archive processing: Analyzing archives of the world wide web. In *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2002)*, pages 16-31, Rome, Italy, September 16-18 2002. Springer.  
<<http://link.springer.de/link/service/series/0558/papers/2458/24580016.pdf>>.

[25] Royal Library, Sweden. Kulturarw3 - Long time preservation of electronic documents. Website.  
<<http://www.kb.se/kw3>>.

[26] Web Warehousing and Mining Group. WHOWEDA. Website, April 2002.  
<<http://pipe.cais.ntu.edu.sg:8000/~whoweda>>.

[27] T. 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>>.

Copyright © Andreas Rauber, Andreas Aschenbrenner, Oliver Witvoet, Robert M. Bruckner and Max Kaiser

---

Top | Contents  
Search | Author Index | Title Index | Back Issues  
Opinion | Next Article  
Home | E-mail the Editor

---

D-Lib Magazine Access Terms and Conditions

DOI: 10.1045/december2002-rauber