

# Austrian On-Line Archive Processing: Analyzing Archives of the World Wide Web

Andreas Rauber<sup>1</sup>, Andreas Aschenbrenner, Oliver Witvoet

Department of Software Technology, Vienna University of Technology  
Favoritenstr. 9 - 11 / 188, A-1040 Wien, Austria  
<http://www.ifs.tuwien.ac.at>

**Abstract.** With the popularity of the World Wide Web and the recognition of its worthiness of being archived we find numerous projects aiming at creating large-scale repositories containing excerpts and snapshots of Web data. Interfaces are being created that allow users to surf through time, analyzing the evolution of Web pages, or retrieving information using search interfaces. Yet, with the timeline and metadata available in such a Web archive, additional analyzes that go beyond mere information exploration, become possible. In this paper we present the AOLAP project building a Data Warehouse of such a Web archive, allowing its analysis and exploration from different points of view using OLAP technologies. Specifically, technological aspects such as operating systems and Web servers used, geographic location, and Web technology such as the use of file types, forms or scripting languages, may be used to infer e.g. technology maturation or impact.

**Keywords:** Web Archiving, Data Warehouse (DWH), On-Line Analytical Processing (OLAP), Technology Evaluation, Digital Cultural Heritage

## 1 Introduction

In the last few years we have witnessed the initiation of numerous projects aiming at the creation of archives of the World Wide Web. Snapshots 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, characteristics styles of how Web sites were designed to attract visitors, and many other facets of this medium and society in general. Thus, these archives may well end up forming one of the most fascinating collections of popular digital cultural heritage in the future. While several initiatives are already building Web archives [1, 8, 11, 16], several significant challenges remain to be solved, requiring models for preserving the digital artifacts [6], or concepts for cost-efficient distributed storage [5], to name just a few. When it comes to the usage (or prospected usage, as many of these archives currently provide limited or no access to their collections, out of legal or technical reasons), most projects focus solely

---

<sup>1</sup> Part of this work was done while the author was an ERCIM Research Fellow at IEI, Consiglio Nazionale delle Ricerche (CNR), Pisa, Italy.

on the content-aspect of their archive. Interfaces are developed that allow users to surf through time, see the evaluation of a Web page from one crawl to the next, or trace the growth of the Web space of a given site.

Yet, with such a repository of Web data, as well as the meta-data 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 a variety of technical platforms as well as data representations to make its information available. The providers of information are located in different physical places on the hyperlinked 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 where information was being produced and replicated, which technology was used for representing a certain kind of information, what kind of systems were used to make the information available. It also gives us the means to trace the life cycle of technology, following file formats, interaction standards, and server technology from their creation, via different degrees of acceptance to either prolonged utilization or early obsolescence. It provides a basis for tracking the technological evolution of different geographical areas, analyzing characteristics such as the “digital divide”, not from a consumer’s point of view, i.e. who has access to Web information, and who has not, but also from a provider’s point of view, i.e. which areas in the world, as well as on a much smaller, regional scale, are able to make themselves heard, are able to participate in the exchange of information by publishing information on their own account on the Web.

The answers to these kind of questions require a different perspective of the Web and Web archives, focusing not solely on content, but on the wealth of information automatically associated with each object on the Web, such as its file format, its size and the recentness of its last update, its link structure and connectivity to other pages within the same site, domain, and externally, the language used, the operating system and Web server software running on the server side machine, the physical location of the machine, the use of specific protocols, cookies, and many more.

We address these issues in 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. In order to support these kind of analyzes in a flexible manner, we adopt a solution based on a Data Warehouse for the Austrian On-Line Archive Processing module (AO-LAP), allowing interactive analysis of the accumulated data using on-line analytical processing techniques. On top of the Data Warehouse, additional Data Mining techniques may be applied to analyze and characterize specific problem domains, such as time-series prediction for technological evolution.

The remainder of this paper is organized as follows: Section 2 provides an overview of related work in the field of Web archiving, navigation and analysis. This is followed by a presentation of the Austrian On-Line Archive (AOLA) in Section 3. The principles of Data Warehouse technology and OLAP processing are

briefly introduced in Section 4, followed by a description of the current AOLAP system in Section 5. Section 6 gives initial results. We finally provide an outlook on future work in Section 7.

## 2 Related Work

In the last years we have witnessed the creation of numerous initiatives building archives of the World Wide Web. Among the most famous of these we find, for example, the *Internet Archive* [10, 11], located in the US, which, among many other collections, has the largest archive of Web pages from all over the world, donated by the search engine Alexa. Within Europe, the leading project with respect to Web archiving is the *Kulturaw3* project by the Swedish Royal National Library [1]. Its archive contains frequent snapshots of the Swedish national Web space starting in 1996, using the *Combine* harvester as their means of data acquisition. A second large initiative in this field is the archiving initiative of the *NEDLIB* project [8, 18], headed by the Finish National Library and the Helsinki Center for Scientific Computing. Within the scope of the project, a special crawler specifically geared towards tasks of Web page archiving, has been developed, and is currently being used to acquire a snapshot of the Finish Web space. This tool has also been used by other national groups, e.g. in Iceland, to build collections of their respective Web space. Similar initiatives are being followed e.g. in the Czech Republic by the National Library at Brno, the National Libraries of Norway and Estonia, and others.

With respect to the usage of these Web archives, the *Nordic Web Archive* initiative [14] is currently developing an access interface, that will allow users to search and surf within such an archive. A similar interface, called the *Wayback Machine*, is already available for the *Internet Archive*, providing, for each URL entered, a timeline listing the dates when this specific URL was added to the archive, i.e. which versions of the respective file are available.

Going beyond the mere navigation within the archive as a mirror of the World Wide Web existing at the respective times, several projects take a more structured approach to storing and analyzing the Web. The *Web Archaeology* project [13] studies the content of the World Wide Web using a variety of content representations, referred to as features, including *links* capturing connectivity, *shingleprints* capturing syntactic similarities, and *term vectors* capturing semantic similarities. The *Mercator Extensible Web Crawler* is used for large-scale data acquisition, and specific database models were developed at the second layer of the system architecture for storing the feature databases. Various tools are added to the top layer of the system architecture to facilitate specific types of analysis, such as, e.g. in the *Geodesy* project trying to discover and measure the structure of the Web.

Another Web page repository is being built within the *WebBase* project at Stanford University, addressing issues such as the functional design, storage management, as well as indexing modules for Web repositories [9]. The main goal of this project is to acquire and store locally a subset of a given Web space in order to facilitate the performant execution of several types of analyzes and queries, such as page ranking, and information retrieval. However, it limits its scope to the archiving

of one copy of each page at a time, thus providing no historization, and focuses on html pages only.

On a different level we find the *WHOWEDA* project, pursued by the Web Warehousing and Data Mining Group at the Nanyang Technological University in Singapore [2]. Within this project, Data Warehouse technology is used for the storage of consecutive versions of Web pages, adding a time dimension to the analysis of content and link structure. URL, size, date of last modification (and validity, with respect to subsequent visits to a given site), size, etc. are stored together with the content and structure of a document. Furthermore, link information, as well as the position of links within documents are recorded and made available for further analysis. Although a more structured approach to the analysis of Web pages is taken within the scope of this project, it primarily focuses on a detailed analysis and representation of the content of the documents.

### 3 AOL: The Austrian On-Line Archive

The *Austrian On-Line Archive*<sup>1</sup> (AOLA) [17,16] is an initiative to create a permanent archive documenting the rise of the Austrian Internet, capturing the sociological and cultural aspects of the Austrian Web space. The Austrian Web space covers the whole *.at* domain, but also servers located in Austria yet registered under “foreign” domains like *.com*, *.org*, *.cc*, etc. are included. The location of these servers so far is determined semi-automatically by a WHOIS lookup for links to non-at domains encountered and subsequential addition to a list of allowed servers. Furthermore, sites dedicated to topics of Austrian interest as well as sites about Austria (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.aci.org> are examples for such sites of interest. These sites are fed into the system using a currently manually maintained list.

Web crawlers, specifically the Combine Crawler, are used to gather the data from the Web. While the crawling process itself runs completely automatically, manual supervision and intervention is required in some cases when faulty URLs are encountered. The pages downloaded from the Web are stored together with additional metadata in a hierarchical structure defined by the Sweden’s *Kulturaw3*-project, and archived in compressed format on tapes.

The data and the associated metadata gathered from the crawl by the AOLA project are the basis for our analysis within the AOLAP project. The archive currently consists of about 488 GB of data from two crawls, with more than 2,8 million pages from about 45.000 sites from the first partial crawl in 2001 (118 GB in total), as well as about 370 GB (approx. 8,2 Mio URLs from more than 120.000 individual servers, which amount to about 170.000 different servers including alias names of servers) from the second crawl in spring 2002.

---

<sup>1</sup> <http://www.ifs.tuwien.ac.at/~aola>

## 4 Data Warehousing and OLAP

When it comes to the analysis of large amounts of data in a flexible manner, Data Warehouses (DWH) have evolved into the core components of Decision Support Systems. In this section we will briefly sketch the main characteristics of DWHs in general, without being able to address issues of DWH design and different types of data models used for subsequent analytical processing in detail. We rather refer to the wealth of literature on DWH design for these issues, e.g. [12, 15].

A Data Warehouse is a subject-oriented, integrated, time-variant, non-volatile collection of data in support of decision-making processes. 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. One of the core components of any DWH analysis is, contrary to conventional transaction-oriented database systems, the core functionality of the time dimension, facilitating analysis of the development of data across time periods. To achieve this, data, rather than being up-dated or deleted, is only added to a DWH with reference to a validity time-stamp (or rather: a range of time-stamps, depending on the concept of time used, such as valid time, revealed time, etc. See [3] for a detailed treatise of time-related aspects in DWH maintenance).

The main concept of a DWH is the separation of information into two main categories, referred to as *facts* and *dimensions*, respectively. Facts is the information that is to be analyzed, with respect to its dimensions, often reflecting business perspectives, such as a geographic location, evolution over time, product groups, merchandising campaigns, or stock maintenance. No matter whether the data is actually stored in a flat relational DBMS using a dimensional design, such as the star or snowflake models, or whether a multi-dimensional DBMS is used, the DWH may be viewed as a multi-dimensional data cube. This data cube allows us, using OLAP tools, to interactively drill-down, roll-up, slice and dice, to view and analyze the data from different perspectives, derive ratios and compute measures across many dimensions. The *drill-down* operation can be used for example to navigate from the top-level domains to the sub-level domains. The inverse *roll-up*, when applied to the aggregation of total links from hosts which are located, e.g., in Graz (a city) may result in an aggregation of links from hosts located in Styria (the respective county). The *slice* operation defines a sub-cube by performing a selection on, for instance, *domain = .ac.at* on the dimension *domains*, to get all information concerning the educational Internet domain in Austria. The *dice* operation defines a sub-cube by performing selections on several dimensions. For example, a sub-cube can be derived by interactively dicing the cube on three dimensions resulting in the generated clause, *county = "Vienna" and operating system = "linux" and web-server = "apache"*.

These OLAP operations assist in interactive and quick retrieval of 2D and 3D cross-tables and chart-table data from the cube which allow quick querying and analysis of a Web-linkage data storage.

## 5 AOLAP: Austrian On-Line Archive Processing

In this section we outline the various types and sources of information used in the context of the AOLAP project, comment on some feature extraction and transformation steps, as well as on the design of the DWH.

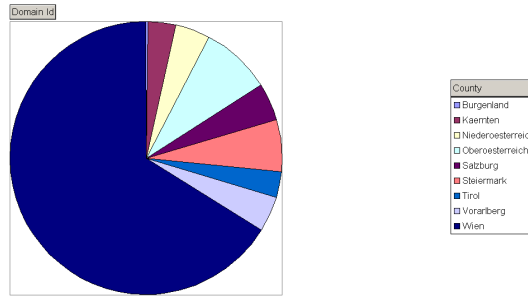
As the primary source of information we use the data gathered by the *Austrian On-Line Archive (AOLA)* project. The archive consists of Web pages, including all types of files as collected by the harvesting software, and rests on tape archives organized primarily according to domain names. In addition to the actual pages, meta-information that is provided or created during the crawling process, is collected and 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, such as the server software type and version, the operating system used by the server, date and time settings at the server, as well as last-modified dates for the respective file being downloaded. This information is stored together with each individual file, encapsulated in MIME format.

Based on this archive, a set of perl-scripts is used to extract a set of relevant data from the files, producing intermediary files that are used for data cleansing and further preprocessing. 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 *frames*, *e-mail addresses* and interactive *forms* used in the case of html files, *date of last modification*, and others. With respect to 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. Please note, however, that the domain name registry information obtained this way, while providing the location of the owner of a specific domain, does not necessarily reflect the actual physical location of a server. Other approaches to evaluate the geographical location may be used, such as directly using host or domain name information, or analyzing references to locations in the textual content [7]. Yet, they do not provide as detailed information, or actually address a different concept, such as in the latter case, a content-based geographic coverage of a site, rather than its location. As we will see during the discussion of the experiments, the inclusion of this kind of content-based geographical coverage, even if it is somewhat less precise, might prove beneficial for detailed analysis.

The information is further transformed and loaded into a relational DBMS using a star-model like design for the data storage. The data model basically consists of two parts. The first part arises from all tables containing data referring to the Web hosts the data derives from. The second part consists of the tables containing data about the hosts where links point to. Connecting these parts is the table where all the links are stored. This table forms the central fact table in the Data Warehouse. Below we provide a brief description of the main tables in the database.

- *domains*: This table contains the names of the Web hosts organized by sublevel domains. This allows us to drill down through the address space during analysis. We also check if the domain is reachable over the Internet at a specific time and store this information in this table. Identical hosts, i.e. hosts under the same IP address, yet reachable via different domain names, are stored multiple times in this table to reflect the actual domain name space.
- *IPs*: The IP addresses of the Web hosts, separated into the octets ranging from 0 to 255 are stored in this table. This is used to identify the different types of IP nets, i.e. class A, B and C networks. Class A addresses are reserved for very large networks such as the ARPANET and other national wide area networks. Class B addresses are allocated to organizations that operate networks likely to contain more than 255 computers and Class C addresses are allocated to all other network operators. Here, servers reachable via multiple domain names are stored only once.
- *server*: In this table server information like the type and version of the server (e.g. MS IIS, Vers 5.0) is stored, structured hierarchically by Producer, Product, and Version. Please note, that no checking of the validity of this information is performed during the download, i.e. disguised servers are not identified as such.
- *OS*: The table OS contains the reported name and the specificity of the operating system of the hosts. Again, no checking of the validity of this information is performed.
- *maintainers, netnames, and netblocks*: These tables contain information gathered from the WHOIS servers. The owner of the netblock in which the specific host is addressed can be retrieved from the maintainer table. In the table netnames, the name of the netblock is stored, and the table netblocks contains the range of the IP addresses of a specific netblock.
- *owner and address*: These two tables as well are filled with data from the WHOIS server. In the first one the owner of the Web host and in the second one the address registered at the WHOIS server, is stored, structured hierarchically by zip code, city and county.
- *pages*: In this table all pages gathered from the AOLA database are stored. There is a column *page* where the name of the page is stored, a column *url* containing the URL of the specific page. Further information includes the size of the page, crawl date, as well as the date of the last modification for the downloaded page if provided by the server.
- *link\_domains*: This table stores the Web hosts which are not in the Austrian Web space (so-called foreign Web hosts) but are linked by the Austrian hosts. The sublevel domains are stored separately as in the table domains described above.
- *link\_pages*: This table contains all the so-called foreign link pages, i.e., pages referenced by pages of the AOLA database which themselves are not in the Austrian Web space, and thus not part of the AOLA archive.
- *forms*: This table stores, for pages in the AOLA archive, the number of forms per page and the total amount of fields of a specific page to facilitate analysis of interactive Web pages, types and amount of interaction encountered etc.



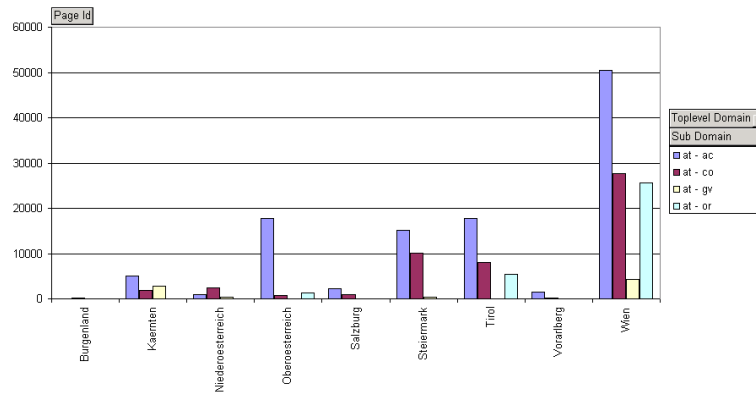
**Fig. 1.** Distribution of Web servers in Austria

- *filetype*: This table contains the different file types of the pages in the archive as well as those of the foreign pages. The information is structured hierarchically by *media*, i.e., application, video, or image, *MIME type* and the *filename extension*. As basis for this structuring, both the MIME type provided with the downloaded page, as well as the file extension are used, forming two independent dimensions. This separation is necessary due to the fact that the information provided both by the MIME type as well as by the file extensions is prone to errors, and quite frequently these two dimensions do not correspond to each other. Retaining both information domains thus provides greater flexibility in the analysis.
- *run*: In order to be able to compare the characteristics of the Austrian Web over time, we have to compare data from different crawls. For each crawl we define a run number, start and end date, stored in the run table.
- *domain\_to\_domain\_links*: All the links we gathered are stored in this table. In the column *type* there are the different prefixes of the URL which indicate the protocol (http, https, ftp, etc.) of the stored link. *External* is an additional column which is used to differentiate between external and internal links, i.e., if the link references a page from another Web host or from the same domain. In the Data Warehouse this table forms the central fact table.

Based on these tables, a multi-dimensional cube is created which can further be used for interactive analysis.

## 6 Experimental Results

In this section we present examples of the analytical capabilities of the AOLAP system. We should emphasize, however, that the current results are based on an incomplete crawl of the Austrian Web space, representing data from the first pilot crawl in spring 2001 and a second crawl started in spring 2002. Thus, the numbers provided below can only depict a trend, rather than be taken as confirmed results yet. However, the large amount of data already available at least allows us to analyze the current situation of the Austrian Web space, as well as obtain ideas of its usage, challenges with respect to its preservation, as well as to discover the benefits of



**Fig. 2.** Distribution of Web servers across counties and domains

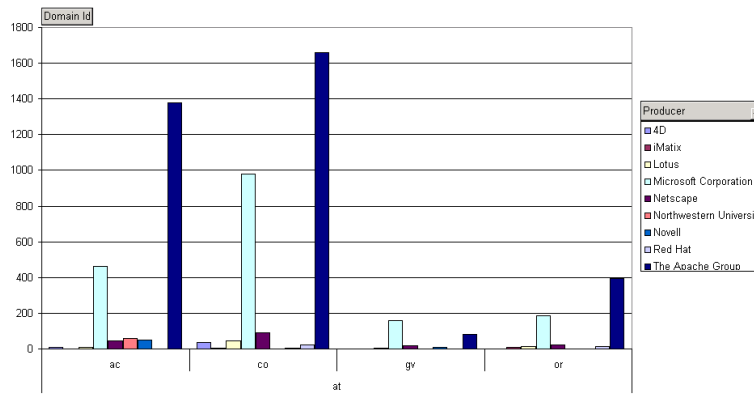
interactive analysis provided by a DWH-based approach. In order to exploit the most important characteristic of such a Web archive, i.e. to analyze its historic perspective and use this as a basis for impact evaluation and trend analysis, a series of snapshots over several years will need to be accumulated in order to facilitate evaluation along the time dimension.

### 6.1 Distribution of Web-servers over the counties

Figure 1 represents a distribution graph of the domains in Austria, showing that most of the Web servers are located in the capital Vienna. If we make another slice by restricting the IP addresses to class A IPs only, the difference is even more obvious. Although this fact is not really surprising, the magnitude of the difference between the metropolis and the rest of Austria still is astounding, especially when we consider that just less than a quarter of the population lives in Vienna. More precisely, our analysis reveals that 66% of the Web-hosts are registered in Vienna, followed by Upper Austria with 9% and Styria with 6%. The distribution of the Web hosts in these other counties are comparable to the distribution of the population. This points towards the much-discussed issue of the metropolitan media Internet.

However, care must be taken with respect to the information represented by the geographical domain, which reflects the location of owner of a certain IP segment, rather than the actual location and area serviced by a specific server. As many nation-wide operating ISPs are based in Vienna, and thus have their address block registered there, the actual saturation and distribution of Internet services is somewhat distorted. A combination with other means of location or geographical coverage determination should be incorporated to cover these issues, such as content-based coverage identification mentioned in Section 5 [7].

A drill-down onto the sub-domains provides a different view of the national distribution, where, for example, the academic and commercial nets are more evenly dispersed among the counties, whereas governmental websites as well as organi-



**Fig. 3.** Distribution of Web servers across domains

zational sites are less wide-spread. Furthermore, we may not forget to take into account the “foreign” hosts, i.e. hosts registered in Austria, but registered under foreign domains, which currently amount to more than 6.500 individual domains (or close to 9.000 if alias names of servers are considered independently).

## 6.2 Distribution of Web servers across domains

While the distribution of the different Web servers used on the Web is one of the most frequently analyzed facts, and thus in itself does not reveal any surprising results, the application of DWH technology allows us to more flexibly view the various facets of this subject. We came across 35 different types of servers or server producers, in a total of about 300 different versions, but the most common ones are the APACHE and the IIS server, followed by the Netscape-Enterprise server. For a selection of the various types encountered, see Table `reftab:servertypes`.

By drilling-down we can take a look at the distribution of Web servers at the first sub-domain level. Figure 3 depicts the resulting distribution focusing on the most prominent types of Web servers. The general trends in market shares remain more or less unchanged, with probably a slightly higher presence of the Apache Web server in the academic domain. However, an interesting characteristic is represented by the presence of the WN Web server from Northwestern University, an open-source Web server that basically is used exclusively in the academic domain. This difference becomes even more obvious when we take a look at the relative distributions, depicted in Figure 4. Here the dominance of Apache is clearly visible in all domains, but further characteristic differences can be detected, such as the more homogeneous spreading across a variety of servers in the organizational domain. Another characteristic easily detectable is the large number of different servers employed in the commercial domain, as opposed to the more limited spectrum in the governmental domain.

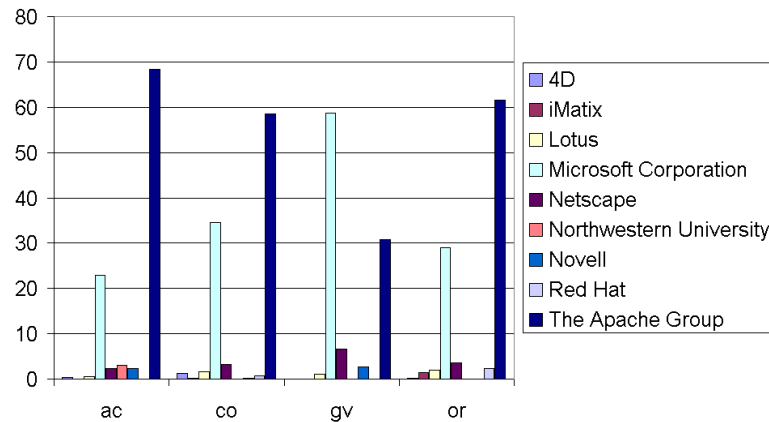


Fig. 4. Relative distribution of Web servers across domains

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

The number of file types encountered in the Web archive is of high relevance with respect to the preservation of the archive, in order to keep the pages viewable in the near and far future. It also represents a good mirror of the diversity of the Web with respect to the technologies employed for conveying information. All over 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 criterium. However, we should probably stress, that the quality of the information provided this way is very low, as a large number of both file extensions as well as MIME types are actually invalid, such as files with extensions *.htmo*, *.chtml* or *.median*, *.documentation*. A listing of some of the most important types of files found in the archive is provided in Table 2. For a comprehensive overview of almost 7.000 different file extensions and their associated applications, see [4] While the major part of file extensions encountered definitely are erroneous, they point towards 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.

Several interesting aspects can be discovered when analyzing the distribution of file types across the different types of Web servers. General known tendencies, like the dominance of the PDF format over the previously very important Postscript file format for document exchange can be verified this way, as depicted in Figure 6.

Figure 5 depicts the distribution of various types of video file formats across Web servers. Here we find significant differences 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 *Quick-time*, which is less than half as popular, but still ahead of various other video formats identified by their MIME type as flavours of *ms-video*. (We also find a video format identified as MIME type

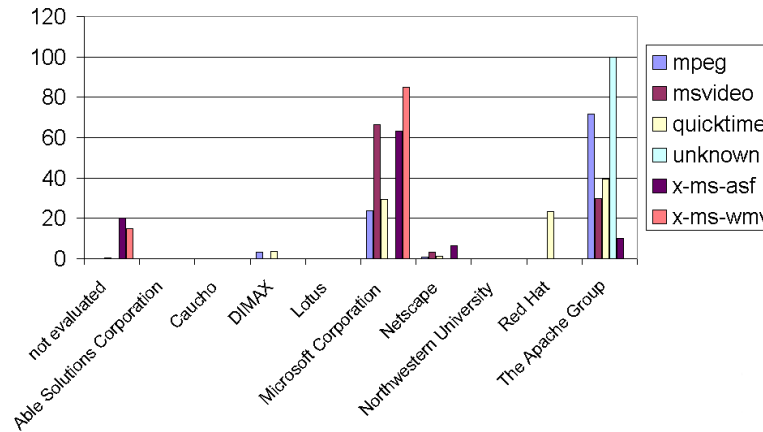
Producer	# Versions	# Occurrences
4D	13	361
Able Solutions	1	10
Apple ShareIP	5	24
Caucho	5	19
DIMAX Hyperwave-Info-Server	4	68
IBM	14	78
Lotus Lotus-Domino	2	506
Microsoft IIS	7	20947
NCSA	4	48
Netscape	26	1509
Northwestern Univ.	1	63
Novell	3	138
RapidSite	2	438
Red Hat Stronghold	6	297
Roxen	2	102
The Apache Group	52	47383

**Table 1.** Selection of server types and versions encountered

*video/unknown* on Apache servers. By viewing 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 sharply contrasted by the situation encountered at Web sites running the MS IIS Web server, where the latter family of formats by far dominates the type of video files provided. 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 *Quick-time* movies. Untypical distributions like this 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, that in this case the distribution can be attributed to a sub-group of 10 domains out of several hundred sites using the Stronghold server. Of these 10 sites, 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.

Similar characteristics can be detected when analyzing image file type distributions across different server types as depicted in Figure 7. Here we find an almost exclusive presence of the *png* filetype on Apache servers, whereas almost 60% of all *bmp* files are to be found on MS IIS servers. However, when we take a look at the absolute distributions, we find that the *png* file format still plays a neglectable role at this time, with a clear dominance of *jpeg*, followed by *gif* images.



**Fig. 5.** Distribution of video file types across Web servers

## 7 Conclusions

We have presented the Austrian On-Line Archive Processing (AOLAP) system providing a flexible means to analyze data stored in a Web archive. The main benefit of the proposed approach lies in the flexibility with which interactive analysis may be performed. Contrary to most current approaches, the focus of this type of analysis is not primarily on the content of the data, but rather on meta-information about the data as well as about the technologies used to provide a given service. While the improvement of Web search results may be facilitated by the collection and integration of additional information such as link structure analysis, by far more fascinating insights into the Web and its evolution will become possible, providing a basis for crucial technology decision. These include the evolution and maturation of technologies employed, analysis of market shares, but also, from a preservation perspective, technologies and efforts required to preserve the diversity of information representation. While several of these issues have been addressed in various projects employing special purpose tools, the integration of the wealth of data associated with the Web into a Data Warehouse opens the doors for more flexible analysis of this medium.

We are currently taking a look at further types of information to be extracted from the pages, integrating 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, etc., in order to be able to incorporate future technologies. Furthermore, the addition of a content-based dimension is being considered. As part of these expansions flexible interfaces to modify/increase the number and type of technologies to be scanned for in the data, will be analyzed. Furthermore, the application of specific data mining techniques for specific problem domains, especially with respect to time-line analysis will be studied in greater detail.

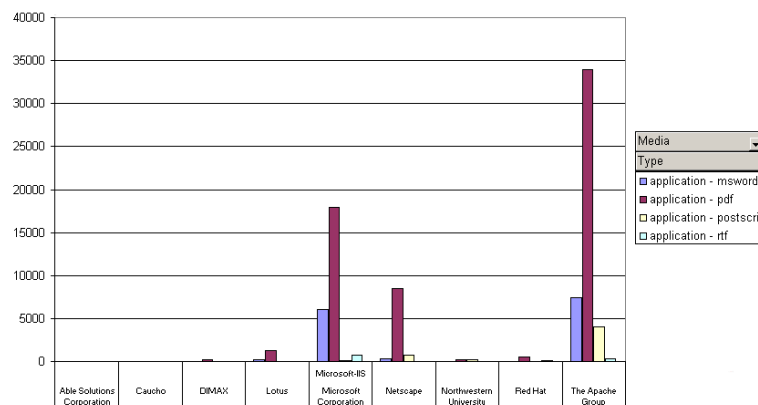


Fig. 6. Distribution of document file types across Web servers

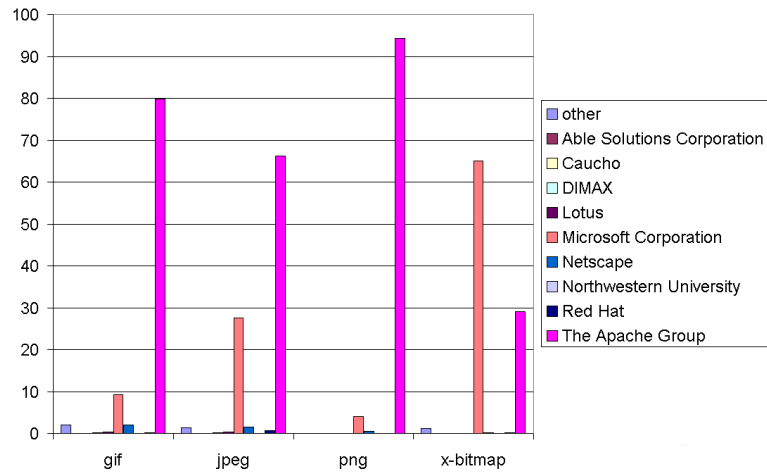
## References

1. 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>.
2. S. Bhowmick, N. Keong, and S. Madria. Web schemas in WHOWEDA. In *Proceedings of the ACM 3rd International Workshop on Data Warehousing and OLAP*, Washington, DC, November 10 2000. ACM.
3. R. Bruckner and A. Tjoa. Managing time consistency for active data warehouse environments. In *Proceedings of the Third International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2001)*, LNCS 2114, pages 254–263, Munich, Germany, September 2001. Springer. <http://link.springer.de/link/service/series/0558/papers/2114/21140219.p%df>.
4. Computer Knowledge (CKNOW). FILExt: The file extension source. Webpage, June 2002. <http://filext.com/>.
5. 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>.
6. M. Day. Metadata for digital preservation: A review of recent developments. In *Proceedings of the 5. 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.
7. J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of web resources. In *Proceedings of the 26th International Conference on Very Large Databases, VLDB 2000*, pages 545–556, Cairo, Egypt, September 10-14 2000.
8. 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>.
9. 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*

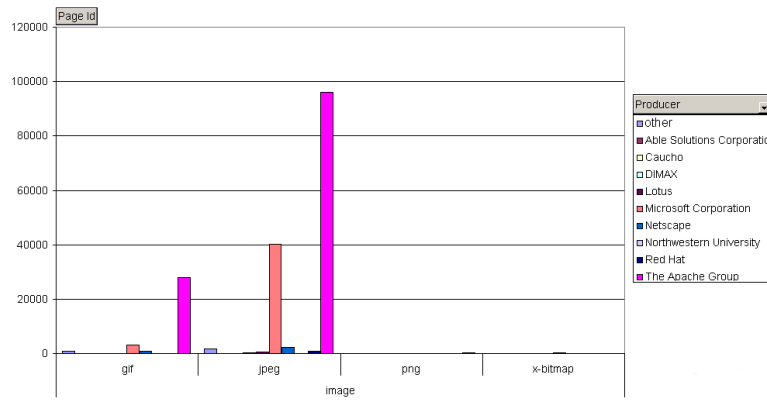
MIME type	# Occ.	MIME type	# Occ.
Application/ms-excel	1227	Image/gif	35144
Application/ms-powerpoint	841	Image/jpeg	145200
Application/msword	14799	Image/png	349
Application/octet-stream	9916	Image/tiff	1025
Application/pdf	67976	Image/x-bitmap	426
Application/postscript	5274	Image/other	123
Application/x-dvi	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

**Table 2.** Selection of MIME types encountered

- (*WWW9*), Amsterdam, The Netherlands, May 15-19 2000. Elsevir Science. <http://www9.org/w9cdrom/296/296.html>.
10. The Internet Archive. Website. <http://www.archive.org>.
  11. B. Kahle. Preserving the internet. *Scientific American*, March 1997. <http://www.sciam.com/0397issue/0397kahle.html>.
  12. R. Kimball. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. John Wiley & Sons, 2 edition, 2002.
  13. 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>.
  14. Nordic web archive. Website. <http://nwa.nb.no>.
  15. T. Pedersen and C. Jensen. Multidimensional database technology. *IEEE Computer*, 34(12):40–46, December 2001.
  16. A. Rauber. Austrian on-line archive: Current status and next steps.
  17. A. Rauber and A. Aschenbrenner. Part of our culture is born digital - On efforts to preserve it for future generations. *TRANS. On-line Journal for Cultural Studies (Internet-Zeitschrift für Kulturwissenschaften)*, 10, July 2001. <http://www.inst.at/trans/10Nr/inhalt10.htm>.
  18. 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>.



**Fig. 7.** Relative distribution of image file types across Web servers



**Fig. 8.** Absolute distribution of image file types across Web servers